INSTITUTO SUPERIOR DE EDUCAÇÃO E CIÊNCIAS

ESCOLA DE COMUNICAÇÃO ARTES E INDÚSTRIAS CREATIVAS

PÓS-GRADUAÇÃO EM GESTÃO E COMUNICAÇÃO DE UM PROJETO EM DESENHO APLICADO – MENTION IN RESEARCH

Curso 2023- 2024

ESCOLA DE COMUNICAÇÃO, ARTES E INDÚSTRIAS CRIATIVAS ISEC LISBOA

Censored Bodies: The Impact of Algorithmic Moderation on Sex-Positive and Queer Identities

Carly Sheridan

carly.n.sheridan@gmail.com

Orientadores:

Rafael Pozo Puértolas (PhD)

Miguel Jaime (PhD)

Em colaboração com:

tgraf isec lisboa

CENTRO INTERNACIONAL DE ESTUDOS E INVESTIGAÇÃO EM TECNOLOGIAS GRÁFICAS E COMUNICAÇÃO CIENTÍFICA

Acknowledgements

I would like to thank Andrés Colmenares (BA, MA), Director of the Master's Degree in Design for Responsible Artificial Intelligence (Elisava Barcelona School of Engineering and Design); and Rafael Pozo Puértolas (PhD) and Miguel Jaime (PhD), Coordinators of the Postgraduate Degree Mention in Research (ISEC Lisboa and Elisava).

My sincere thanks to faculty members Caroline Sinders and Eryk Salvaggio for their thoughtful discussions and academic support of this project as it developed. I'd also like to recognize my learning peer, Ela Pušica, for being a sounding board.

I would like to express deep gratitude for all those who contributed insights and data to this project, either through interviews or as survey respondents. Sharing these narratives helped frame my research and provided a much needed, deeply human lens. Lastly, I want to acknowledge my family for their support and encouragement, specifically my father, Patrick, and my partner, Leonardo.

Abstract

The problem identified in this study is the pervasive and discriminatory censorship of sex-positive and marginalized bodies in online spaces. It addresses the digital censorship caused by content classification systems and binary dataset designs. It examines the categorizations used by artificial intelligence models and algorithms to censor and suppress these bodies. These issues are detected predominantly on social media platforms where algorithmic and human-driven moderation disproportionately targets content from sex workers, sex-positive educators, and individuals identifying as particular sexualities and gender identities. This research contributes to the discourse of the necessity of designing inclusive platforms and tools. The proposed solution is a critical and novel redesign of the traditional database and advocates for data labeling as a form of empowerment. It challenges the consolidation of moral authority by major technology companies, questioning why we collect the data points that we do while advocating for a more contextual and temporal approach to data.

Methodologically, the research combines qualitative interviews with content analysis, focusing on the lived experiences of sex-positive individuals impacted by digital censorship. Through these interviews, the study illustrated the nuanced ways in which algorithmic biases and societal standards work together to silence voices, hinder economic opportunities, and impede on efforts to find and build community online. The results highlight that censorship is not only about content removal, but also carries implications for identity, visibility, and self-expression. The proposed solution is important because it seeks to address the broader implications of such censorship, which include the misrepresentation and erasure of diverse bodies, and the negative impact of how sex and sexuality are perceived, discussed, and learned.

Keywords

Censorship; sexuality; sex-positivity; inclusion; dataset design, algorithmic moderation

Supplementary index

(i) Glossary

Algorithm	A step-by-step procedure for solving a problem or performing a task.
Algospeak	The adaptation of language through code words or alternate spelling to evade algorithmic censorship.
Artificial Intelligence	In simplified terms, artificial intelligence (AI) is a computer's ability to mimic human intelligence.
Bodies	In the context of this paper, bodies refer to the physical, embodied structure of a human being and the social and cultural constructions around the human body.
Censorship	The suppression or prohibition of speech or access to platforms on the basis of its content being considered "mature", harmful, violent, or sexually explicit.
Content regulation	The way information on the internet is controlled, imposed by laws, policies, practices, and people.
Data ethics	The branch of ethics that studies and evaluates the moral implication of data collection, storage, processing, and usage, particularly personal or sensitive data. Privacy, consent, fairness, transparency, and security are key issues data ethics considers when evaluating the potential harms or benefits.
Data labeling	The process of assigning a value or more information to a data point so that machine learning models can more easily

	sort, reference, and make predictions from the data.
Dataset	A dataset is a socially constructed collection of information. Datasets can be compiled by a single person or by a company.
Dataset design	The process of creating, curating, and designing, datasets used to train machine learning models.
Deviant	As an adjective, departing from usual or accepted standards, especially in social or sexual behavior. As a noun, a deviant person or thing.
Inclusion	The practice of actively ensuring all people, regardless of background, gender, race, sexual orientation, and body type are welcomed, respected, and able to fully participate in digital spaces.
Machine Learning	Machine learning (ML) is a branch of AI that uses data and algorithms to make predictions.
Sextech	Technology that is designed to enhance and innovate human sexuality and the human sexual experience.
Shadow banning	A form of censorship where content is fully or partially blocked, but the creator is unaware of the censorship.

(ii) List of acronyms and abbreviations

AI	Artificial Intelligence
et al.	and others (Latin: et alii)
DD	Deviant Dataset All interviewees have been anonymized using DD and the numerical order in which they were interviewed
DEI	Diversity, Equity, and Inclusion
LGBTQIA+	Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, Asexual, and more
ML	Machine Learning
NSFW	Not Safe For Work
FOSTA/SESTA	FOSTA (Allow States and Victims to Fight Online Sex Trafficking Act) SESTA (Stop Enabling Sex Traffickers Act)
VSD	Value Sensitive Design

Index

A3 Guiding Principles

A3.1 Design Justice Network Principles

A3.2 Privacy by Design

A3.3 Data Feminism Principles

“Missing data sets” are the blank spots that exist in spaces that are otherwise data-saturated. That which we ignore reveals more than what we give our attention to. It’s in these things that we find cultural and colloquial hints of what is deemed important. Spots that we've left blank reveal our hidden social biases and indifferences.

Mimi Ọnụọha
The Library of Missing Datasets

1. Introduction

1.1 Study Area

Any time spent in the physical and natural world is an opportunity to expand one's worldview. Whether through food, language, fashion, or culture, we are exposed to bodies that look, sound, and exist differently than our own. As we collectively spend more time in virtual spaces – the average internet user spends six hours and 35 minutes online daily[1] – and with the proliferation of artificial intelligence and algorithmic decision-making, the bodies we are exposed to on the internet are increasingly homogenous. Heavily filtered, both aesthetically and through content moderation, these images can in turn shape our perception of the world. What happens when your body is not reflected back to you on the screen? This unremitting provocation planted the seed that ultimately led to the research presented here today.

The motivation behind this research is rooted in the concern over the consolidation of power within a few major technology companies, which impacts the representation and visibility of diverse bodies and experiences. The reduction of human experiences to mere data points, optimized for machine readability, strips away the nuances of human sexuality and identity. While categorization is necessary for AI models, the nuanced nature of human sexuality and its ethical representation requires more sophisticated and inclusive approaches. Sexuality is not only central to the human experience, it is the foundation of our lineage, what connects us to our ancestors and future generations; it is an exploration of our deepest desires and our most sacred selves. This study aims to address the ethical implications of such reductions, particularly focusing on how sex-positive and queer bodies are disproportionately censored online.

[1] Average daily time spent using the internet by online users worldwide from 3rd quarter 2015 to 4th quarter 2023, Statista

This research is also inspired by a belief that when we design platforms and tools for our most marginalized and those historically excluded, we will build online spaces that are thus safe for everyone.

The researcher's capacity to carry out this work stems from a combination of academic knowledge, professional experience, and a deep personal commitment to addressing the ethical challenges posed by modern technologies. Through the Master in Design for Responsible AI program at Elisava Barcelona School of Design and Engineering, the researcher has honed their skills in creative research, context analysis, and critical thinking. It has provided a comprehensive framework for investigating the multifaceted impacts of AI systems on daily life and for addressing complex questions related to digital technologies in sustainability, ethics, and social justice from intersectional and transdisciplinary perspectives.

As a white, straight, cisgender woman, the researcher is keenly aware of the privilege they embody and the responsibilities that come with it. They recognize that while those impacted by flawed and harmful systems should be integral to addressing these issues, it is unjust to expect the oppressed to bear the burden of fixing the broken systems they did not create. Thus, the researcher uses their positionality as a central component of their ability to carry out this research and to advocate for greater inclusivity in tech, ensuring that efforts to dismantle these oppressive structures are collaborative and supportive rather than exploitative. This research represents a first attempt on the researcher's behalf at creating a solution to these critical issues, informed by a commitment to ethical and inclusive technology.

1.2 Experimentation

This research is titled "Censored Bodies: The Impact of Algorithmic Moderation on Sex-Positive and Queer Identities." It investigates the systemic biases within AI-driven content moderation systems and their impact on the visibility of marginalized bodies.

To explore these issues, this study employs a multifaceted methodology. The research includes a comprehensive literature review to define existing biases in AI and content moderation systems, alongside qualitative ethnographic interviews with individuals affected by the censorship of these technologies. Additionally, an analysis of social media platforms' content moderation policies and practices provides insight into the discriminatory nature of automated censorship.

Together, this mixed-method approach allows for a thorough examination of the intersection between AI, gender identities, sexual identities, and body representation. The literature review establishes a theoretical backdrop, the interviews offer personal and contextual insights, and the platform policy analysis grounds the research in current digital practices. This methodological triangulation ensures a comprehensive exploration of the research questions from multiple angles, enhancing the validity and depth of the study's findings.

1.3 Contribution

This research advances the understanding of how AI intersects with gender and sexual identities, contributing to the discourse on sex-positivity. It contributes to the discourse on AI ethics and digital inclusivity by providing theoretical insights and practical frameworks for creating more inclusive AI systems. Central to this research is the development of the Deviant Dataset, a novel, yet speculative, dataset design proposal. Drawing from principles of Design Justice, Privacy by Design, and Data Feminism, the Deviant Dataset challenges traditional data practices and promotes more holistic, inclusive methodologies that can be integrated into AI development.

The findings of this research further validate how current AI practices reinforce societal biases, particularly against non-normative gender identities and bodies. By advocating for the development of more contextual and individualized data collection methods, this study provides a framework for more ethical and equitable technologies. Furthermore,

the research highlights the need for policies that protect marginalized communities through implementations like the right to be forgotten and the importance of self-identifying input options over standard checkboxes. It proposes that digital platforms adopt more nuanced and temporal strategies, moving beyond binary classifications and embracing the complexity of human identities.

2. Problematic

2.1 Contextual Framework

The investigation is situated within the broader domain of database systems and machine learning, focusing particularly on how these technologies intersect with social constructs such as gender and sexuality. This research moves from a general examination of database structures and data classification, exploring the technical frameworks and societal implications, to a specific focus on how these systems perpetuate biases and exclude what has been deemed non-normative identities. By analyzing the historical and contemporary contexts of database design and machine classification, the study aims to uncover the ways in which these technological practices reinforce existing social inequities.

The study area is defined as the intersection of database structures, data classification systems, and their sociocultural impacts, particularly concerning gender and sexuality. This involves a detailed examination of the ways in which databases encode human attributes, the historical context of binary gender classification, and the ongoing challenges posed by these rigid structures in a society that increasingly recognizes gender as a spectrum. The investigation also includes an analysis of the implications of these systems for marginalized communities, such as nonbinary, queer, trans, intersex, and gender-nonconforming individuals, and how legislative measures like FOSTA/SESTA–which will be expanded upon– further complicate the landscape for sex workers and their representation in digital spaces.

The study subjects include databases, machine learning models, and the individuals and communities affected by these systems. Databases are not merely technical artifacts but are embedded with social values that reflect and perpetuate societal norms. The interaction between these systems and marginalized communities reveals significant tensions, as databases often fail to accommodate diverse identities, leading

to misrepresentation and exclusion. This misalignment is particularly evident in the binary classification of gender, which is rooted in outdated sociocultural contexts and fails to capture the complexities of contemporary understandings of identity. The study also considers the role of sex workers as a marginalized group significantly impacted by the interplay of technological design and legislative actions, highlighting the broader implications for online discourse and community support.

The researcher's position in this study is both analytical and advocative. Methodologically, the researcher employed qualitative interviews and content analysis to explore the lived experiences of sex-positive individuals affected by digital censorship. The researcher positions themselves as a critical observer, acknowledging the sociotechnical nature of the systems under study. This involves a commitment to examining both the technical details and the broader social implications of database structures and classification systems. By adopting a sociotechnical perspective, the researcher aims to illuminate the ways in which technical designs are influenced by and reinforce social values and power dynamics. This position also entails a critical engagement with the ethical dimensions of technology design, emphasizing the need for more inclusive and flexible systems that better reflect the diversity of human experiences, diverse sexualities, and identities.

2.1.1 Graphic Representation of the Contextual Framework

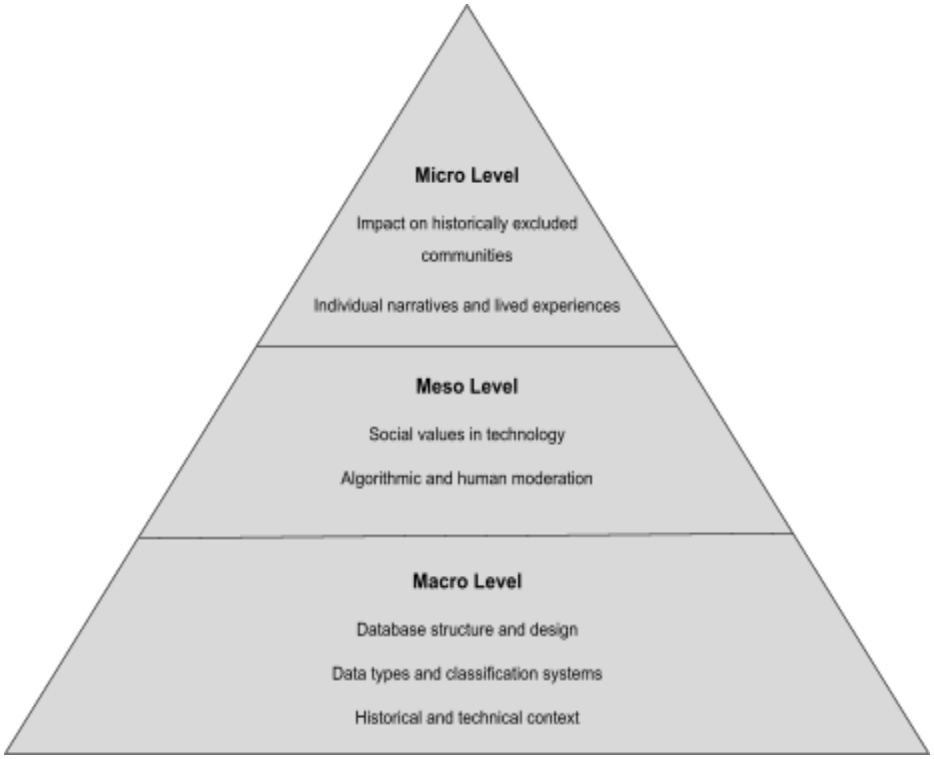
The graphic representation of the contextual framework can be visualized as follows:

1. Macro Level (research area)
    a. Databases and Machine Learning
        i. Structure and design
        ii. Data types and classification systems
        iii. Historical and technical context

2. Meso Level (sociocultural context)
    a. Social Values in Technology
        i. Gender and sexuality encoding
        ii. Sociocultural norms and biases
        iii. Historical roots of binary classification
    b. Algorithmic and Human Moderation
        i. Content classification systems
        ii. Social media platforms' policies

3. Micro Level (study area)
    a. Impact on Historically Excluded Communities
        i. Nonbinary, queer, trans, and gender-nonconforming individuals
        ii. Sex workers, sex educators, sextech founders
        iii. Legislative impacts (FOSTA/SESTA)
        iv. Shifting policies of social media platforms
    b. Individual Narratives and Lived Experiences
        i. Censorship, misrepresentation, and erasure
        ii. Content removal, account suspension, self-censorship, and community-driven tactics

4. Researcher's Position
    a. Analytical Approach
        i. Qualitative interviews
        ii. Content analysis
    b. Critical Engagement
        i. Sociotechnical perspective
        ii. Ethical implications
        iii. Inclusive and flexible design considerations
    c. Advocative Stance
        i. Promoting novel platform design
        ii. Ethical data labeling, collection, and storage practices



The contextual framework is designed to illustrate the layered nature of the investigation, moving from a broad examination of database systems and their technical underpinnings to a focused analysis of their sociocultural impacts. At the macro level, the research area encompasses the fundamental components and principles of databases and machine learning. This sets the stage for deducting how these systems function technically.

At the meso level, the study contextualizes these technical systems within the broader sociocultural environment, examining how social values and norms are embedded in technological design. This includes an analysis of the historical context of gender classification and the ongoing challenges posed by these rigid structures.

The micro level brings the focus to specific communities historically excluded and affected by these systems, highlighting the real-world implications of technical design

choices. By examining the impact on nonbinary, queer, trans, intersex, and gender-nonconforming individuals, as well as sex workers, sextech founders, and sex educators, the study reveals the significant social consequences of database structures and classification systems.

The researcher's observation position is crucial for maintaining a critical and reflective approach, ensuring that both technical and social dimensions are considered. By adopting a sociotechnical perspective, the researcher aims to provide a comprehensive analysis that not only critiques existing systems but also offers insights into more inclusive and equitable design practices. This framework underscores the importance of integrating social understanding into technological solutions, emphasizing that fairness and inclusivity are not merely technical challenges but deeply intertwined with broader societal values and power dynamics.

2.2 Study Area

The study area encompasses the intersection of database structures, classification systems, and sociotechnical systems, with a specific focus on the implications of these systems for marginalized identities. Databases, as essential components of modern computing systems, are instrumental in organizing, storing, and retrieving data. Their design and implementation reflect not only technical considerations but also entrenched social values and norms, which significantly influence the inclusivity and representation within these systems.

At the commencement of this investigation, the “state of the art” in the study area revealed several critical trends and issues. Databases predominantly relied on rigid classification schemes that often failed to accommodate the diversity and fluidity of human identities. For instance, the binary gender classification (Male/Female) persisted in many systems, reflecting outdated sociocultural norms that do not align with contemporary understandings of gender as a spectrum (Broussard, 2023). This rigidity in classification systems has profound implications, particularly for nonbinary, queer, trans, intersex, and gender-nonconforming individuals, whose identities are often invalidated or misrepresented within computational infrastructures.

The study subjects involved in this investigation include the structural elements of databases, the algorithms that govern classification and categorization processes, and the human individuals whose data is being classified and categorized. These subjects interact in complex ways, with the technical components of databases and algorithms often perpetuating societal biases and exclusions. For example, algorithms used in content moderation and biometric security systems frequently misidentify or misclassify individuals based on their characteristics, leading to broader issues of discrimination and marginalization (Katyal, Jung, 2021).

The current state of the problem is characterized by a significant misalignment between social progress and technical implementation. Despite advancements in recognizing

and respecting diverse identities and experiences, many database systems and classification algorithms remain anchored in archaic models that reinforce existing power dynamics and social hierarchies. This disconnect is not merely a technical oversight but a reflection of deeper sociocultural inertia that resists change.

In the context of sex work, the study area examines the impact of digital platforms' policies and practices on sex workers' visibility, safety, and community engagement. Legislative measures like FOSTA/SESTA have exacerbated the deplatforming and shadow banning of sex workers, limiting their access to essential online spaces for communication, support, and advocacy. These issues illustrate the broader challenges of designing inclusive and equitable sociotechnical systems that can adapt to and reflect the complexities of human life.

2.3 Research Problem

The research problem identified in this study area is the pervasive misalignment between the design of database structures and classification systems and the diverse, evolving realities of human identities and experiences. This misalignment manifests in various ways, including the persistence of binary gender classifications, the exclusion of marginalized groups from mainstream digital platforms, and the reinforcement of social biases through algorithmic decision-making processes.

Exploring this problem is crucial for several reasons. First, the technical design of databases and classification systems has far-reaching implications for social justice and equity. When these systems fail to recognize and accommodate diverse identities, they contribute to the marginalization and disenfranchisement of already vulnerable populations. For instance, the use of binary gender classifications in databases not only misrepresents nonbinary individuals but also perpetuates a form of systemic violence that invalidates their identities.

Second, the issue of algorithmic bias and discrimination highlights the need for a more nuanced understanding of how social values and technical design intersect. Algorithms used in content moderation, biometric security, and other applications often reflect and reinforce societal prejudices, leading to discriminatory outcomes. Addressing these biases requires a comprehensive approach that integrates technical, social, and ethical considerations.

Third, the challenges faced by sex workers in navigating digital platforms underscore the broader issues of visibility, safety, and community support in online spaces. The deplatforming and shadow banning of sex workers not only hinder their ability to work safely and communicate effectively but also reinforce stigmatization and isolation. By examining the impact of platform policies and practices on sex workers, this research aims to shed light on the broader implications of digital censorship and the need for more inclusive and supportive digital environments.

By critically examining the intersection of database structures, classification systems, and human identities, this research seeks to contribute to a deeper understanding of how technical design can better reflect and support the complexities of human life. This endeavor is not only a technical challenge but also a social imperative, demanding a re-evaluation of the values and assumptions that underpin our technological infrastructures.

3. Theoretical Framework

3.1 Antecedents and References

The theoretical framework for this research is rooted in the intersection of sociotechnical systems, data justice, and the representation of marginalized identities within digital infrastructures. The antecedents and references provided form a comprehensive foundation that elucidates the multifaceted nature of the research problem. The following sections will discuss the most relevant theoretical arguments and contributions from various scholars in this domain, linking these to our study's focus on database structures, classification systems, and the impacts on gender, sexuality, and sex work.

Sociotechnical Systems and Data Bias

Meredith Broussard's book *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech* (2023) provides a critical examination of how technological systems embed and perpetuate biases against marginalized groups. Broussard argues that these biases are not mere glitches but are indicative of systemic issues within the design and implementation of tech systems. This argument underscores the importance of scrutinizing how databases and classification systems reflect and reinforce societal prejudices, particularly against nonbinary, queer, and gender-nonconforming individuals.

> “Lines of code can change the world, absolutely…Computer systems are not just mathematical. They are sociotechnical, and they need to be extensively updated on a regular basis. Just like humans.” (Broussard, 2023)

Anna Lauren Hoffmann's research further elucidates the intricate relationships between data, discourse, and violence. In *Terms of Inclusion: Data, Discourse, Violence* (2021) and *Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse* (2019), Hoffmann critiques the limitations of current fairness and

antidiscrimination frameworks. Hoffmann posits that these frameworks often fail to address the deeper structural inequalities embedded in data systems, thus perpetuating exclusions and injustices while allowing tech companies to promise to “do better” with very little actual accountability structures in place.

> “Inclusion operationalizes the language of remorse and “doing better” to obscure its commitment to a substantive, technorationalist conception of the future—that is, a future where more and better data science and technology alone can save us from the very problems data science and technology generate.” (Hoffmann, 2021)

Intersectionality and Representation

Patricia Collins' seminal work *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment* (1990) and Kimberlé Crenshaw's foundational essay *Demarginalizing the Intersection of Race and Sex* (1989) provide essential theoretical underpinnings for examining how intersecting identities of race and gender are navigated within sociotechnical systems. Collins and Crenshaw emphasize the importance of intersectionality in understanding the unique challenges faced by individuals who exist at the confluence of multiple marginalized identities.

Safiya Umoja Noble's *Algorithms of Oppression: How Search Engines Reinforce Racism* (2018) expands on these ideas by demonstrating how search engines and similar technologies can perpetuate racial biases through their algorithms. Noble's work highlights the critical need for reevaluating and redesigning these systems to ensure they do not perpetuate existing social inequities. This concept aligns with the broader objective of our research to develop more inclusive and equitable database structures.

French historian and philosopher Michel Foucault's *The History of Sexuality: An Introduction* (1978) offers a crucial historical and philosophical context for understanding how power dynamics shape the discourse around sexuality. Foucault's analysis of the

ways in which societies control and categorize sexual behavior provides a foundational framework for examining the role of database structures in enforcing or challenging normative sexualities. His concept of biopower—wherein power is exercised over bodies and populations through regulatory mechanisms—resonates with the study of how databases classify and control information related to gender and sexuality.

> "Without even having to pronounce the word, modern prudishness was able to ensure that one did not speak of sex, merely through the interplay of prohibitions that referred back to one another: instances of muteness which, by dint of saying nothing, imposed silence. Censorship." (Foucault, 1978)

Digital Platforms and Marginalized Communities

The impact of digital platforms on marginalized communities, particularly sex workers, is another critical area of concern. Amber Davisson and Kiernan Alati's study '*Difficult to Just Exist': Social Media Platform Community Guidelines and the Free Speech Rights of Sex Workers* (2024) investigates how community guidelines on social media platforms affect sex workers' ability to communicate and engage online. Their findings illustrate the significant constraints and risks faced by sex workers due to platform policies, reinforcing the need for more inclusive and supportive digital environments.

> "Certain identities, particularly those of sex workers, tend to be culturally coded as sexual content. A sex worker's very existence on the site, even if they are not posting primarily about their work, can be treated as sexual content." (Davisson & Alati, 2024)

Clare Southerton, Daniel Marshall, Peter Aggleton, Mary Lou Rasmussen and Rob Cover's paper *Restricted Modes: Social Media, Content Classification and LGBTQ Sexual Citizenship* (2021) explore the specific ways in which social media content classification affects LGBTQ communities. Southerton et al. argue that restrictive content classification systems on social media platforms undermine LGBTQ "sexual

citizenship” (Southerton, et al.) by limiting the visibility and expression of non-normative sexualities. This work highlights the pressing need to reform content classification practices to promote a more inclusive digital space for LGBTQIA+ individuals.

Hanne M. Stegeman, Carolina Are, and Thomas Poell's article *Strategic Invisibility: How Creators Manage the Risks and Constraints of Online Hyper(In)Visibility* (2024) provides a contemporary perspective on the strategies used by online content creators to navigate the risks of hypervisibility. Their research highlights the delicate balance creators must maintain to avoid censorship while still engaging with their audience. This study is particularly relevant for understanding how classification systems and platform policies impact the visibility and agency of marginalized content creators.

Algorithmic Fairness and Critical Perspectives

Rob Kitchin's *Thinking Critically about and Researching Algorithms* (2017) offers a methodological framework for analyzing the role of algorithms in sociotechnical systems. Kitchin emphasizes the need for critical approaches to studying algorithms, considering their socio-political contexts and impacts. This perspective is instrumental for our research, which seeks to uncover how classification algorithms in databases affect the representation and inclusion of diverse identities.

> “Algorithms are created for purposes that are often far from neutral: to create value and capital; to nudge behavior and structure preferences in a certain way; and to identify, sort and classify people.” (Kitchin, 2017)

The work of Andrew D. Selbst, Danah Michele Boyd, Sorelle Alaina Friedler, Suresh Venkatasubramanian, and Janet Vertesi in *Fairness and Abstraction in Sociotechnical Systems* (2019) further develops the discourse on algorithmic fairness. They argue that traditional notions of fairness often fall short in addressing the complexities of sociotechnical systems, advocating for more nuanced and context-specific approaches. This aligns with our research objective to move beyond simplistic fairness frameworks

and towards more comprehensive strategies that account for the diverse realities of human identities.

Sonia K. Katyal and Jessica Y. Jung's *The Gender Panopticon: Artificial Intelligence, Gender, and Design Justice* (2021) adds a critical dimension to the discussion by examining how artificial intelligence and gender interact within the framework of design justice. Katyal and Jung argue that AI systems often perpetuate gender biases through their design and implementation, advocating for a justice-oriented approach to AI development. Their insights are crucial for our research, which aims to explore how database structures can be redesigned to promote gender inclusivity and justice.

> “Studying the impact of AI on transgender and nonbinary populations also provides us with a window that helps to assess how all populations are gendered and surveilled, in the marketplace and beyond, as a result.” (Katyal & Jung, 2021)

The antecedents and references provided form a robust theoretical foundation for this research. These works collectively highlight the critical need for rethinking and redesigning database structures and classification systems to better reflect and support the complexities of human identities. By integrating insights from sociotechnical studies, intersectional theory, and critical algorithmic research, this study aims to contribute to the discourse.

3.2 Argument

This research aims to validate the argument that AI systems and algorithms inherently perpetuate biases, leading to the systematic exclusion and stigmatization of sex-positive and marginalized bodies. The study critically examines how these biases distort public discourse on sex and sexuality, reinforcing narrow societal norms and limiting personal expression. By addressing the research problem and main questions, the study seeks to provide empirical evidence on the extent and impact of algorithmic biases, ultimately arguing for the necessity of more inclusive and equitable AI design practices.

3.2.1 A Historical and Systemic Approach to Sexuality

In *The History of Sexuality: An Introduction, Volume 1*, Michel Foucault explores the relationship between power and sexuality, providing a historical context through which we can better comprehend the contemporary issue of algorithmic moderation on sex-positive bodies. His analysis reveals how power structures have historically regulated sexuality, creating frameworks that continue to influence modern digital censorship practices.

Foucault situates the advent of sexual repression in the seventeenth century, marking a departure from previous eras characterized by more open expressions of sexuality. He argues that “silence became the rule” regarding sex, with repression functioning not only as a mechanism of disappearance but also as an “injunction to silence, an affirmation of nonexistence” (Foucault, 1978). This repression aligns with the rise of capitalism, where sexuality became incompatible with the general and intensive work imperative of the time. According to Foucault, the systematic repression of sex served to support the emerging economic and social order by ensuring that sexual discourse was confined to narrow, regulated bounds.

In the eighteenth and nineteenth centuries, this repression evolved. Foucault notes a “political, economic, and technical incitement to talk about sex,” but not in an

emancipatory sense. Instead, sex became a matter for analysis, classification, and control, reflecting the interests of a state increasingly concerned with managing populations (Foucault, 1978).

This period saw the transformation of sexuality into a "police matter," essential for understanding demographic variables such as birth and death rates, fertility, and health. The state claimed it required knowledge of its citizens' sexual practices to maintain social order and optimize productivity. This expansion of control mechanisms underscored a pervasive surveillance network aimed at regulating every aspect of sexual behavior. Foucault highlights how the categorization and pathologization of sexual identities, such as homosexuality, emerged during this period, reflecting a deeper entrenchment of power over the body and its desires.

The concept of power in Foucault's work is not merely repressive but also productive. He argues that “power operated as a mechanism of attraction; it drew out those peculiarities over which it kept watch” (Foucault, 1978). This dual impetus of pleasure and power do not cancel each other out but instead reinforce one another through complex mechanisms of excitation and incitement. The interplay of these dynamics is crucial for understanding the contemporary landscape where digital platforms, through algorithmic moderation, continue to regulate and often suppress sexual expression. Foucault's historical analysis chronicles that the regulation of sexuality is deeply rooted in the mechanisms of power that have evolved over centuries. The “triple edict of taboo, nonexistence, and silence” (Foucault, 1978) that characterized earlier periods persists in modern forms of censorship. These systems, though technologically advanced, echo the historical patterns of control and repression Foucault describes. His work provides a critical lens to examine how today’s digital practices are part of a long continuum of sexual regulation. The historical trajectory he outlines underscores the persistent entanglement of sexuality and power, highlighting the need for this research as an act of ongoing resistance and reconfiguration of these dynamics.

3.2.2 Database Structures and Data Types

Databases, as fundamental components of computer systems, play a crucial role in organizing, storing, managing, and retrieving data. They facilitate the efficient handling of vast amounts of human-sourced data through systematic organization and very precise classification. The structure and design of databases reflect not only technical requirements but also social values, which can have significant implications for representation and what is and is not included in them, particularly in how they encode and categorize human attributes such as gender or sexuality.

A simple database record might be represented as follows (Broussard, 2023):

- Firstname [string]
- Lastname [string]
- Gender (M/F) [Boolean]
- Address 1 [string]
- Address 2 [string]
- Zip [number]

Each field holds data that adheres to a particular type—such as strings for text, numbers for numerical values, and Boolean for binary values, as represented above (Broussard, 2023). This structure ensures consistency and efficiency in data storage and retrieval as each field imposes strict rules on what kind of data can be entered. For instance, in the American Standard Code for Information Interchange (ASCII), the letter 'A' is represented as 01000001, exemplifying the strict and limited nature of data types used in databases. These constraints, while ensuring efficiency and minimizing errors, also reinforce normative aesthetics known as “elegant code.” This concept values maximum speed and efficiency, often at the expense of flexibility and inclusivity. As such, a Boolean value for gender might offer an incremental gain in efficiency but also perpetuates a form of violence against nonbinary, queer, trans, intersex, and gender-nonconforming individuals by invalidating their identities within computational

systems (Broussard, 2023).

3.2.3 Classification and Categorization

Classification within databases is a process where data is organized based on predefined categories or attributes. These categories often reflect societal norms and conventions, which can become problematic when they fail to adapt to evolving understandings of identity. Historically, for example, gender classification in databases has been binary. This design choice is not merely a technical decision but one deeply rooted in the sociocultural context of the 1950s in the United States and United Kingdom, reflecting the era's rigid binary conception of gender as immutable and strictly male or female (Broussard, 2023).

Despite significant advances in understanding gender as a spectrum and recognizing LGBTQIA+ rights, many contemporary systems still adhere to this outdated model. The persistence of binary gender representation in databases illustrates a broader issue within computational infrastructures. The classification systems designed by early computer scientists were influenced by their commitment, whether conscious or unconscious, to maintaining a rigid and retrograde status quo (Broussard, 2023). “Even though sweeping social change happened in the 1960s and the 1970s, including second wave feminism and the civil rights movement and gay rights and the widespread recognition that sex is biological and gender is socially constructed, academic computer science pointedly ignored the topic of gender except to think about how a computer might accurately translate gendered pronouns from one language to another,” she writes.

One prominent example of the limitations in current database design is Facebook's, now named Meta, handling of gender data. Although Meta allows users to self-identify their gender beyond the binary options, the underlying system still records users as male, female, or null. This approach effectively nullifies any gender identity outside the binary,

demonstrating how databases can enforce cisnormativity and erase nonbinary identities (Broussard, 2023).

3.2.3.1 AI and Gender Identities

The intersection of AI and gender presents profound challenges that expose the limitations and biases embedded within current technological and legal frameworks. The reliance on binary gender classifications by AI technologies creates a form of erasure for those who do not conform to these rigid categories. Such technologies, from biometric surveillance to content filters on social media, misrecognize and exclude transgender and nonbinary populations, leading to both practical and existential harms. These harms include the denial of rights and services, the amplification of stigma and discrimination, and the economic disadvantages stemming from content demonetization and visibility issues on platforms like YouTube (Katyal & Jung, 2021).

AI technologies, particularly those involved in surveillance and categorization, operate on a binary understanding of gender that fails to accommodate the diversity of gender identities, leading to significant issues for transgender and nonbinary individuals. Katyal and Jung draw on the concept of the panopticon, originally articulated by Jeremy Bentham, to critique the pervasive and invasive nature of AI surveillance in the context of gender and sexuality. Bentham's panopticon, designed to exert control through constant visibility, is mirrored in modern AI systems that discipline and regulate gender expression through perpetual surveillance (Katyal & Jung, 2021).The pervasive misrecognition of gender identity by AI systems thus functions as a mechanism of social control, reinforcing normative gender expectations and marginalizing non-conforming individuals.

AI's failure to accurately recognize and respect diverse gender identities is rooted in the biases inherent in its training data and the design choices made by predominantly non-diverse developers (Katyal & Jung, 2021). This results in datasets that reflect and

perpetuate existing societal biases, producing discriminatory outcomes. For instance, AI systems often classify transgender and nonbinary individuals as errors, outliers, or deviants furthering their invisibility and subjecting them to heightened scrutiny and surveillance (Katyal & Jung, 2021). This systemic bias not only undermines the rights of these individuals but also perpetuates a cycle of exclusion and discrimination.
The concept of “gender reduction” services, as opposed to “gender recognition” services, aptly describes the reductive nature of AI's approach to gender. These technologies simplify complex identities into binary categories, which fails to catalog the fluidity and multiplicity of gender experiences. This reductionist approach leads to the overrepresentation of normative gender identities and the erasure of non-normative ones, thus reinforcing traditional gender norms and stereotypes (Katyal & Jung, 2021). Moreover, the economic motivations behind these technologies, such as the need to sell data to advertisers, further entrench this binary framework, as evidenced by Facebook's limited pronoun options despite a more diverse range of gender identities (Katyal & Jung, 2021).

The panoptic nature of AI surveillance also extends to social media, where platforms' content moderation policies disproportionately affect LGBTQIA+ communities. These policies often conflate LGBTQIA+ content with sexual explicitness, leading to unjust censorship and further marginalization. The economic and social consequences of such censorship are significant, as they impact the visibility, income, and community engagement of LGBTQIA+ content creators (Katyal & Jung, 2021). This reinforces a digital environment where non-normative identities are constantly monitored, regulated, and suppressed, echoing the controlling mechanisms of the panopticon.

DD4[2], a queer woman, has faced censorship and content removal due to her refusal to conform to these narrow standards. She noted how societal expectations on women's appearances are perpetuated and amplified by AI, which often promotes idealized body images that exclude diverse and authentic representations of women. “As women, we

[2] All interviewee respondents have been anonymized, further explanation can be found under Methodology, Interviews

are taught from a young age to look a certain way. I did not realize that as I got older, AI would be doing the same," she said.

The exclusion of non-normative bodies from digital spaces not only reinforces harmful stereotypes but also perpetuates a culture of invisibility and exclusion. DD4's experiences highlight the psychological toll these biases can have as well, leading her to reduce her presence on social media despite her desire to contribute to the discourse on body positivity and diversity. This reduction in visibility is a direct consequence of the negative feedback loops and censorship, reflecting the broader disenfranchisement felt by many individuals whose bodies and identities are marginalized by algorithmic content moderation systems.

3.2.3.2 Social Media and Sexual Citizenship

How social media platforms like YouTube and Tumblr regulate LGBTQIA+ content through their classification systems imbed normative understandings of "sexual citizenship." As defined by Southerton et al. (2021) in *Restricted Modes: Social Media, Content Classification and LGBTQ Sexual Citizenship,* the concept of sexual citizenship refers to the diverse claims of belonging that individuals make based on their sexual identities and practices. The content regulation that happens on these sites reveals a tension between the inclusive potential of digital spaces and the restrictive mechanisms imposed by content classification practices. YouTube and Tumblr, particularly through their Restricted Mode and Safe Mode features, have implemented changes since 2017 that significantly impact the accessibility of LGBTQIA+ content (Southerton et al., 2021).

These features, intended to filter out 'mature' content, often misclassify LGBTQIA+-related material as inappropriate, effectively censoring it. These systems rely heavily on algorithmic sorting, supplemented by human moderators, to categorize content. However, the processes are often opaque, with platforms offering minimal transparency about how decisions are made. Users are typically informed only of the

final decision, which is presented as objective despite the proprietary nature of the algorithms used (Southerton et al., 2021).

The authors argue that these platforms function as “norm-producing technologies” (Southerton et al., 2021), where the complexities of queer sexuality and desire are often obscured to promote a sanitized version of LGBTQIA+ identities that conform to heteronormative expectations. This process effectively marginalizes more diverse and authentic expressions of LGBTQIA+ sexual citizenship, according to the authors.

Algorithmic classifications have been critiqued for reinforcing existing social prejudices, as seen in Noble’s (2018) work on algorithmic oppression, which argues that discriminatory outcomes are a fundamental feature of these systems. On platforms like YouTube and Tumblr, such classification practices often blur the lines between sex and sexuality, leading to the unwarranted censorship of non-explicit LGBTQIA+ content. This is evident in cases where non-pornographic depictions of LGBTQIA+ life are flagged and restricted as if they were explicit, illustrating the failure of these systems to distinguish between different forms of representation (Southerton et al., 2021).

Moreover, the enforcement of these classification guidelines often involves significant human intervention, which is usually downplayed by the platforms. Human moderators, working under challenging conditions, are expected to make nuanced decisions about content within extremely short time frames. This human element introduces variability and cultural biases into the classification process, undermining claims of objectivity and consistency.The restrictive classification regimes on YouTube and Tumblr not only reflect but also reinforce a normative framework of sexual citizenship. This framework promotes a ‘good’ LGBTQIA+ citizen who adheres to heteronormative standards of acceptability, often excluding more diverse and radical expressions of queer identity. As the platforms navigate the boundaries of what constitutes acceptable LGBTQIA+ content, they perpetuate a sanitized version of queer life that aligns with mainstream, heteronormative values (Southerton et al., 2021).

Despite the restrictive nature of these classification systems, their high-profile failures provide opportunities for users to challenge and resist the imposed norms. By exposing the inconsistencies and biases in content classification, users can advocate for more nuanced and context-sensitive approaches to online governance. These failures highlight the fluidity and complexity of sexual and gender identities, pushing for a re-evaluation of how content is classified and regulated (Southerton et al., 2021).

Take, for example, the experience of DD2[3]. When DD2 posted a photo from a professional boudoir photoshoot and the image was immediately taken down, she knew her account was being treated differently. Censorship has become intertwined with her existence online, whether by the platforms she uses or by her own monitoring determined by the regularity of these instances. Putting into words the experience of having one's body censored online is complex, especially because a sense of self transcends physical appearance. When content is flagged and removed for simply showing skin, regardless of context, it can feel like an attack on one’s identity.

Working as a lingerie model and content creator on platforms like Instagram, TikTok, and OnlyFans, she is subject to an invisible yet oppressive evaluation layer online. “It's not even necessarily tied to anything nude or suggestive, it's the moment I show skin,” she explained. DD2 exists in a precarious place. As a cisgender, straight person, she is aware of her privilege. As a plus-sized woman and person of color who promotes body positivity and capitalizes economically off of her body, she embodies a unique intersectionality that magnifies her experiences of discrimination and marginalization. According to Kimberlé Crenshaw's framework of intersectionality (Crenshaw, 1989), these overlapping identities do not merely add layers of bias, but intersect to create a form of oppression that is more complex than the sum of its parts. For DD2, this means that the societal pressures and prejudices she faces are compounded by her gender, race, and body size, creating a multifaceted experience of exclusion and censorship. These algorithmic biases and societal standards—ones she refuses to

[3] All interviewee respondents have been anonymized, further explanation can be found under Methodology, Interviews

accept—collectively silence her voice, reduce her visibility, and ultimately hinder her ability to fully express and monetize her authentic self online. It also prevents her from sharing often empowering content with other women who might find confidence in seeing bodies like theirs being represented.

“How I show up online and how I show up in person is very much living in inclusion,” she said. “I don’t even post photoshoots anymore, I’ve completely censored my own Instagram. On TikTok, I can't use certain hashtags like #photoshoot. Sometimes my videos won't even get posted, they will immediately go under review before being able to post it.”

Content is removed under the guise of violating guidelines, yet the guidelines themselves are often vague and discriminatory in practice (Davisson & Alati, 2024). When the mere presence of certain bodies triggers these automated censorship systems, it suggests that certain bodies are inherently more sexual and, therefore, inappropriate. Automated systems that flag and remove content without human oversight are often making arbitrary decisions that feel personal, stifling creative expression and creating an “uneven playing field” as DD2 describes it. This censorship can have deeper psychological ramifications as well, instilling a sense of shame and self-consciousness that can hinder a person’s sense of sexuality and self-worth.

The message is insidious and it applies to a growing list of bodies. LGBTQIA+ bodies are labeled as “mature content” (Katyal & Jung, 2021) on a number of platforms, reducing their visibility and reinforcing harmful stereotypes about non-heteronormative identities. This technological censorship reflects broader societal prejudices, where bodies that do not conform to the dominant cultural ideals—that is those that deviate from the thin, white, able-bodied standard put in place—are more likely to be policed and silenced. This censorship operates within what Patricia Hill Collins describes as the "matrix of domination," (Collins, 1990) where multiple axes of power and oppression intersect. For sex-positive bodies, particularly those of marginalized groups, the matrix of domination manifests in a digital environment that systematically privileges certain

types of bodies over others, even in spaces that ostensibly promote freedom of expression.

Censorship of sex-positive bodies is not just about controlling content; it is about controlling the narrative of who is allowed to be seen and celebrated. It reinforces a culture of exclusion where marginalized bodies are continuously policed and invalidated, and it hinders broader societal acceptance and understanding of diverse bodies and identities.

3.2.4 Datasets as Sociotechnical Systems

Datasets are collections of data often used to train, test, and validate machine learning models and algorithms, which are sets of rules or procedures for solving problems, learning patterns, and making decisions based on the data stored (Sarker, 2021). In the context of content moderation, algorithms are used to filter, prioritize, and sometimes censor information.

But databases are not merely neutral repositories of data; they are deeply influenced by the social and cultural contexts in which they are created. The sociotechnical nature of computational systems underscores the interplay between social values and technical design. This sociotechnical nature means there must be more importance placed on a requirement for these systems to evolve to better reflect the complexities of human life. As Broussard writes, “We superimpose human social values onto a mathematical system…Computer systems are not just mathematical. They are sociotechnical, and they need to be extensively updated on a regular basis. Just like humans.”

The rigidity of existing systems often leads to broader issues, not only tied to gender and sexuality. This misalignment between social progress and technical implementation can be seen in various sectors. Biometric security measures like full-body scanners misidentifying individuals based on characteristics like Black women's hair, Sikh

turbans, or prostheses, further demonstrate the inadequacies of poorly designed systems (Broussard, 2023).

3.2.5 Neutrality and AI

While AI technologies and platforms are often looked at as universal tools deemed fit for a globalized world and to be applied in a variety of circumstances, they are not a one size fits all solution. While algorithms are typically described as statistical biases rather than moral ones, these statistical biases lead to discriminatory outcomes nonetheless (Stinson, 2022). “Biased data sets can also be the downstream result of a different kind of systemic discrimination. That facial recognition algorithms are an order of magnitude less accurate for Black female faces than for white male faces has been attributed to the lack of Black and female faces among the training examples used to build facial recognition systems,” writes Stinson, highlighting that statistical and moral biases are often interdependent. This interdependence demonstrates that ostensibly neutral technologies can perpetuate systemic discrimination.

These findings have been corroborated in a number of studies including in the seminal work, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* (2018). They found the poorest accuracy rates among Black females within facial recognition technologies. Darker-skinned females experienced error rates up to 34 percent higher than their lighter-skinned male counterparts. Discriminatory law enforcement practices, exemplified by the overrepresentation of Black Americans in mugshot data for example, create a vicious time loop of deeply ingrained racism and racial biases within these technologies.

The effectiveness of AI models hinges on the quality of the data they are trained on. However, this data is not immune to biases and distortions. The variety of sources used for training can result in a narrow, skewed perspective. Flawed data sampling, coupled with subjective classifications, introduces a layer of subjectivity that may differ across

individuals or cultures. Once data is extracted and ordered into training sets, it becomes an “epistemic foundation” according to Crawford (2021). As she writes in *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, “Artificial intelligence is not an objective, universal, or neutral computational technique that makes determinations without human direction. Its systems are embedded in social, political, cultural, and economic worlds, shaped by humans, institutions, and imperatives that determine what they do and how they do it.”

The subsequent classification of this data becomes a critical juncture, as it frames how AI systems perceive and categorize the world. AI systems reflect the intersectionality of structural, disciplinary, hegemonic, and interpersonal domains, contributing to the reinforcement of existing power structures and biases (Collins, 1990). These classifications, however, operate under the guise of scientific neutrality.

Structural Domain Laws and policies	***Disciplinary Domain*** How laws and policies are enforced
Hegemonic Domain Culture and media circulation	***Interpersonal Domain*** Individual experiences of oppression

Matrix of domination, Collins (1990)

3.2.6 Sex, Stigma, and AI

The enactment of FOSTA (Allow States and Victims to Fight Online Sex Trafficking Act) and SESTA (Stop Enabling Sex Traffickers Act) in the United States significantly altered the landscape of online discourse, particularly for sex workers, by introducing stringent limits on the free speech protections originally provided under Section 230 of the 1996

Communications Decency Act[4]. These legislative measures have led to the shutdown of numerous online forums and the systematic deplatforming and shadow banning of sex workers on various social media platforms. As a result, sex workers have lost critical online spaces that provided not only safety and professional communication channels but also essential community support and advocacy opportunities, particularly for those from marginalized backgrounds who often face physical isolation and lack support networks (Davisson & Alati, 2024).

Social media platforms have become indispensable tools for sex workers to share health and safety information, seek legal advice, and build community with other professionals. These platforms also allow sex workers to engage in public discourse, challenge stigmas, and advocate for their rights. When the content regulations imposed by these platforms conflate nudity and sexual content, as they are prone to do, sex workers are restricted in their ability to communicate openly and safely. Platforms like Facebook, Instagram, and TikTok impose stringent bans on nudity, often without nuanced distinctions that could accommodate the professional and advocacy needs of sex workers (Davisson & Alati, 2024). “Certain identities, particularly those of sex workers, tend to be culturally coded as sexual content. A sex worker’s very existence on the site, even if they are not posting primarily about their work, can be treated as sexual content,” write Davisson and Alati.

The regulations imposed by these platforms often conflate nudity and sexual content in ways that reflect prevailing US attitudes, thereby restricting sex workers' ability to engage in public discourse and support networks (Davisson & Alati, 2024). Social media guidelines, shaped by platform creators' values, disproportionately impact stigmatized groups by censoring content deemed sexual or obscene, which often includes sex workers' presence and professional activities.

[4] H.R. 1865 - Allow States and Victims to Fight Online Sex Trafficking Act of 2017, United States Congress

The algorithmic moderation practices, such as those on Instagram, further exacerbate these issues by removing content based on flawed AI detection methods (Davisson & Alati, 2024).The shifting policies of platforms like Tumblr, Snapchat, and X (formerly Twitter) further complicate the situation. (Davisson & Alati, 2024). Tumblr's oscillation between permissive and restrictive policies on nudity and sexually explicit content exemplifies the precariousness faced by sex workers who rely on these platforms. The lack of consistency in community guidelines makes it difficult for sex workers to navigate these spaces and build stable online communities.

During the course of this research, two popular sites updated their policies. Etsy, a popular e-commerce site for independently owned companies, announced a ban on “mature” products, highlighting the fluctuating boundaries of what is deemed appropriate content online. Under Etsy’s new policy, sex toys will be banned, as will the sale of vintage adult magazines, while the sale of other sex-related items like handcuffs and harnesses will be permitted (Testa, 2024).

X (formerly Twitter) announced plans to allow NSFW (Not Safe For Work) communities to label themselves as such, preventing their content from being auto-filtered (Perez, 2024). This change acknowledges the significant presence of adult content on the platform and provides a structured way to manage it without outright banning it. According to internal documents, 13 percent of all posts contained NSFW content and it is one of the fastest growing genres on the platform (Perez, 2024).

These recent policy changes reflect the complex relationship with what is deemed adult conduct and ever-changing standards of “appropriateness”, illustrating both the broader challenges of moderating sexual content, balancing community safety with inclusivity, as well as the ongoing struggle of sex-positive businesses against censorship and stigmatization.

Deplatforming and shadow banning also have significant financial implications for independent sex workers. The loss of visibility and access to potential clients directly

affects their livelihoods. These enforcement mechanisms disproportionately impact marginalized groups, including Black performers, fat individuals, and members of the LGBTQIA+ community, further entrenching existing social inequalities (Davisson & Alati, 2024). The relegation of sex work to designated sites, away from mainstream social media platforms, not only isolates sex workers but also reinforces the stigma that sex work must be hidden. This segregation limits public discourse on sex work, reducing opportunities for broader societal understanding and acceptance around sex work but also sexuality at large. “Moving all sex work onto designated sites, away from more mainstream social media platforms, comes with its own problems. It sends a message that sex work needs to be hidden, and it allows economic forces to control conversations about what sex looks like by deciding how different forms of work are categorized, whose work is prioritized, and what bodies are most desirable for sex work,” write Davisson & Alati (2024).

DD5, a sex worker offering both virtual and in-person services, echoes these statements. “I’m not ashamed of my work but the internet, and much of society at large, tries telling me that I should be,” she says. Sex work inherently involves a significant education component, revealing a crucial gap in basic sexual knowledge among adults. “A lot of work as a sex worker is education, whether directly or indirectly,” says DD5. “I teach grown men about anatomy, consent, pleasure, and communication. It’s shocking how little some of them know or understand, even about their own bodies.”

With the pervasive stigma surrounding sex, the dynamics of sexual intimacy are often fraught with discomfort and insecurity, leading many individuals–mostly men–to seek the services of sex workers rather than engage openly with their partners, according to DD5. She says many clients over the years have expressed an ability to discuss or explore desires and curiosities within relationships due to fear of judgment or shame. This underscores the broader societal issues surrounding sexuality and communication, where the stigma and lack of proper education inhibit genuine and open discussions.

Consequently, sex workers frequently become not just service providers but confidants and educators, offering a non-judgmental space for clients to express and explore their sexual identities and preferences without fear of condemnation.

3.2.7 The Threat of Visibility

Recent research also explores the concept of “strategic invisibility” (Stegeman, Are, Poell, 2024) deployed among online content creators, particularly those involved in the creation of sexual content, as a risk management technique in response to pervasive platform surveillance, moderation practices, and harassment. This can manifest in several forms, such as concealing body parts, engaging in private shows, or cultivating smaller, supportive online communities. These tactics aim not only to shield creators from unwanted scrutiny but also to create spaces where they can feel safer and more empowered. The regulatory environment, particularly linked to FOSTA/SESTA as discussed above, have intensified censorship, leading to over-moderation, account deletions, and restrictions on expressive freedoms for creators. This compels creators to adapt their visibility strategies continually, often resorting to self-censorship—as expressed by DD2—or modifying content presentation to comply with shifting platform guidelines (Stegeman, Are, Poell, 2024).

Stegeman, Are and Poell’s study illustrates how diverse this visibility management can be, shaped by intersecting factors such as identity, platform policies, and audience expectations. Creators' ability to exert agency in navigating these dynamics highlights the nuanced ways in which they negotiate visibility to assert control over their online presence while mitigating potential risks (Stegeman, Are, Poell, 2024).

Their research also highlights the importance of experiential learning in developing effective visibility management strategies that align their visibility practices with their goals and values (Stegeman, Are, Poell, 2024). Creators often rely on trial-and-error approaches, learning from their interactions with platforms and audiences to refine their

visibility tactics over time. “Sexual(ity) content creators demonstrate that platform visibility logics can be bent and redirected. These creators seek out visibility to the right audiences while trying to avoid the harm associated with being a sexual person online,” the authors write. By targeting specific audiences while avoiding unwanted attention, creators aim to cultivate supportive and lucrative fan bases while mitigating the risks associated with hyper(in)visibility online. This selective approach not only enhances creators' safety but also contributes to the sustainability of their online ventures (Stegeman, Are, Poell, 2024).

The pervasive presence of “algospeak”—the adaptation of language to evade algorithmic censorship—reflects the broader struggle against digital suppression of certain topics. Requiring individuals to use alternative spellings, or using emojis in lieu of words at all, alters their communication as an attempt to thwart censorship and flagging while still connecting with an audience. DD5 says this kind of constant self-editing can be exhausting and yet also exemplifies a resilient spirit. “Some days I’m so sick of algospeak, I’m sick of having to spell sex as seggs, porn as corn, or use code words or emojis or just edit myself at all,” she says. “On other days, it feels good to know we’ll always fight back, we’ll always find a way. We aren’t going anywhere and it’s on them to keep up with us.”

Users devising workarounds to circumvent platform censorship is a global phenomenon, showcasing a form of resistance within the confines of algorithmically mediated environments. Chinese youth on social media tend to avoid using terms common to the global LGBTQIA+ discourse, finding them increasingly irrelevant amidst evolving gender and sexual cultures or because such terminology triggers internet censorship (Wang & Spronk 2023). This dual process highlights the nuanced interplay between conformity and resistance in the digital negotiation of sexual identity (Wang & Spronk 2023) but extends to the platforms themselves as well. Despite massive market success, platforms like Blued and Aloha in China rebranded to avoid censorship as “the world's leading interest-based, social, and health education network” (Blued) and “man-to-man dating app” (Aloha) (Wang & Spronk 2023).

3.2.8 Economic Impact on Founders

The censorship of sex-positive bodies and companies manifests a profound economic impact on founders and their ventures as well. Discussions surrounding sexuality have encountered substantial resistance in several geographic locations, including Brazil, where traditional media platforms like television and radio have historically treated the topic with derision and moral judgment. This pervasive attitude stifles open discourse, discouraging individuals from understanding or discussing sexuality freely and extends itself to online platforms as well. Consequently, founders in this field face formidable obstacles in promoting their content and businesses.

Sextech companies, aiming to promote inclusive and supportive dialogues about sexuality, have found themselves systematically excluded from mainstream advertising channels. For instance, efforts to utilize Google AdWords or Facebook ads are frequently thwarted as accounts are denied outright, stymieing outreach efforts from the outset. Founders, such as DD3, have had to navigate these restrictions by educating the media and demonstrating that discussions about sex can be engaging, welcoming, and comprehensive while fighting censorship online. DD3 is partner and director of a Brazilian company that brings together some of the largest adult apps in Latin America. “We lost a YouTube account with almost 100,000 followers, where we published weekly videos without any nudity simply because of the topic,” she says. “We've also lost nearly 10 Instagram accounts, even while following all the guidelines.”

The systemic barriers that sex-positive founders encounter limit their capacity to reach and support their audience and add an additional, incredibly time consuming and emotionally taxing, layer of work and operational costs. Frustratingly, DD3 knows that the technological infrastructure exists to more effectively segment and manage content and advertisements, yet the initiative and interest from decision-makers are conspicuously lacking. This inaction further perpetuates stereotypes and prejudices, silencing voices that advocate for sexual diversity and expression. “We know that not talking about sex only reinforces stereotypes and prejudices and supports those who

oppress others who can't express themselves," says DD3. "Therefore, talking about sex involves public health, self-esteem, and giving voice to diversity."

DD1, an American sex and intimacy coach and educator, shares these views from her unique vantage point. "I'm in Silicon Valley and it's a bunch of arrogant men who are really excited about what they can make the machine do," she says. "As a technologist, before you get all hopped up on the features that you can implement, you really have to think of the social, psychological, economic, spiritual, ethical questions which I don't think they are."

Both DD1 and DD3 believe that developers and decision-makers who recognize and support the efforts of sextech companies are poised to spearhead critical changes and maintain a competitive edge as their success is intertwined with broader societal progress and a deeper understanding of the importance of discourse on sexuality.

3.2.9 Critical Contexts

Designing truly fair algorithms and AI systems involves navigating a complex landscape of social, technical, and ethical challenges. The quest for fairness is fraught with difficulties, as these systems are often deeply embedded within and influenced by broader societal structures and power dynamics. Several key critical contexts must be considered to distinguish the multifaceted nature of this endeavor and to develop effective strategies for addressing inherent biases and achieving equitable outcomes.

3.2.9.1 Unattainable Fairness

Some recent discourse in fairness-aware machine learning (fair-ML) is focused on engineering algorithms and models that inherently incorporate fairness as a systemic property of black box systems. This approach has been criticized for missing the broader social and contextual factors necessary for achieving genuinely fair outcomes

or even comprehending the concept of fairness itself (Selbst, 2019). Fairness and justice are intrinsic properties of social and legal systems, not proprieties of the technical tools themselves, so attributing in isolation (Selbst, 2019) and attributing the to technical tools in isolation from their social context represents a category error, or more precisely, as Selbst et al. refer to it, an abstraction error.

In the paper *Fairness and Abstraction in Sociotechnical Systems*, the authors identify five conceptual traps—namely the Framing Trap, Portability Trap, Formalism Trap, Ripple Effect Trap, and Solutionism Trap—that emerge from neglecting the interplay between social context and technology. Each trap illustrates the necessity of integrating social understanding into technological solutions to resolve issues effectively. The Framing Trap, for instance, highlights the limitations of defining fairness within the algorithmic frame, which is primarily concerned with optimizing the relationship between data representations and outcomes. The Portability Trap is rooted in the culture of computer science, which values the creation of highly abstract, transferable code. This abstraction is prized for its “elegance” and reusability but often leads to designs detached from social contexts, thus undermining fairness. Similarly, the Formalism Trap underscores the challenges in mathematically defining fairness. “Because algorithms “speak math,” the mandate within the fair-ML community has been to mathematically define aspects of the fundamentally vague notions of fairness in society in order to incorporate fairness ideals into machine learning,” write the authors. The Ripple Effect Trap emphasizes the unintended consequences of introducing technology into social systems. Understanding the impact of such interventions requires a comprehensive evaluation of how technology interacts with pre-existing social structures and values. Lastly, the Solutionism Trap critiques the fair-ML field's inherent assumption that technical solutions are always necessary or sufficient. In the authors’ words, “By starting from the technology and working outwards, there is never an opportunity to evaluate whether the technology should be built in the first place.” (Selbst, 2019).

Achieving fairness in these systems necessitates embracing a sociotechnical perspective, recognizing that both social and technical elements must be considered in

any design process. Technologies are often perceived as distinct and tractable, but sociotechnical thinking forces us to consider their interconnectedness with often sprawling, unruly, and ever-evolving social systems. Designing for fairness in this context is challenging because of the complexity and fragility of sociotechnical systems.

How technology is designed and developed is influenced by humans and power dynamics. Putting the task of defining fairness at all on the shoulders of technologists and researchers creates an environment where a handful of people decide which problems, and which social groups, are the most meaningful to consider which further exacerbated imbalanced power dynamic and exertion (Selbst, 2019). These concerns can only be addressed, the authors argue, by taking the needs of people typically underrepresented in these processes seriously “and by understanding the power dynamics that prevent these voices from having influence in society to begin with.”

3.2.9.2 Implicit Biases

The inherent challenge of addressing bias within AI models and classification systems is intricately tied to the pervasive and often invisible nature of human biases, as illustrated by the late Daniel Kahneman in his final book, *Noise: A Flaw in Human Judgment*. Kahneman’s research addresses the inherent challenges in addressing biases, particularly biases that are implicit, sometimes generational, and often invisible. Consequently, the endeavor to create unbiased AI systems is fundamentally paradoxical, as it necessitates confronting biases that are not only elusive but also at times deeply ingrained in families, industries, cultures, and society at large.

Implicit biases are subconscious attitudes or stereotypes that affect our understanding, actions, and decisions, according to Kahneman. Unlike explicit biases, which are overt and conscious, implicit biases operate below the level of conscious awareness, making them more difficult to detect and rectify. Kahneman's exploration of noise—variability in human judgment that leads to inconsistent decision-making—provides a framework for understanding why biases are so difficult to address. He defines noise as “unwanted

variability in judgments that should be identical" and highlights that this variability is often overlooked compared to biases, which are systematic deviations from the norm (Kahneman, 2021).

The interplay between noise and bias complicates the identification and rectification of biases in AI, as the noise in human judgment can mask or distort the presence of underlying biases. Through this lens, even when developers strive for objectivity, they are likely to encode biases into AI models that will continue to perpetuate and amplify inequities. This creates a circular problem: how can we design fair systems when our benchmarks for fairness are inherently flawed? While AI can process vast amounts of data and identify patterns beyond human capability, it lacks the contextual understanding and moral reasoning that humans bring (Selbst, 2019). Addressing these challenges requires a multifaceted, interdisciplinary approach that goes beyond technical fixes. The collaboration of not only technologists and those typically underrepresented in the processes as Selbst et al suggest, but also ethicists, anthropologists, policymakers, and a wider range of professionals who can create frameworks and guidelines to help mitigate the impact of biases.

3.2.9.3 The Inclusion Trap

Inclusion, while often heralded as a progressive and necessary goal in various social and institutional contexts, warrants a critical examination of its underlying dynamics and potential pitfalls. The work of scholars like Anna Lauren Hoffmann highlights the complexities inherent in inclusionary practices. Promoted as a progressive narrative promising continuous improvement and mitigation of past harms, this narrative allows stakeholders across industry, academia, and government to issue apologies and promises to "do better" without fundamentally challenging the social and financial structures underpinning big data, social media, machine learning, and artificial intelligence (Hoffmann, 2021).

The process of including marginalized groups within established frameworks can inadvertently sustain the very exclusions it seeks to eliminate, contributing to what Hoffmann refers to as “data violence.” This form of violence, while predating many of the digital technologies we use today, has been significantly magnified by computational tools and techniques since the late 19th century, according to Hoffmann.

Addressing the expansive violences inherent in datafication requires an untangling and exposing of the processes through which we internalize and perpetuate oppressive assumptions, like the naturalization of sociopolitical constructs including race and gender (Hoffmann, 2021). What Hoffmann argues is that ideals including, but not limited to, data ethics and AI ethics frame ethics as an iterative process, “rather than one of justice or radical social transformation.”

We accept that technology perpetuates racism, sexism, ableism, ageism, and other forms of biases and discrimination. While companies acknowledge these harmful outcomes of data technologies, they suggest that these issues can be resolved through more technology and more inclusive designs. This approach implicitly admits the harms of previous iterations rather than questioning the fundamental premise of certain data collection and technological deployments. As Hoffmann (2021) writes, “It neutralizes critical calls to *not* collect certain kinds of data or build and deploy certain technologies by reframing the issue as exclusively one of iteration, improvement, and doing things more inclusively.” Furthermore, this approach to inclusivity as design and development issues positions tech companies as the solution. This normalization of dependency on technical expertise not only fails to address, but perpetuates, the potential for violence. Datafication demands the transformation of individuals into reductive, computationally friendly typological features, contingent on conformity to predefined types that fit the system (Hoffmann, 2021). This represents what Hoffmann refers to as “the discursive excess of inclusion”, where “inclusion discourses do not simply normalize, but dupe us into celebrating the very power structures that generate asymmetrical vulnerabilities to violence in the first place.”

> “Inclusion represents an ethics of social change that does not upset the social order. Inclusion positions a certain kind of technology (and, more often than not, tech company) as integral to social progress.” (Hoffmann, 2021)

While inclusion efforts may at times have altruistic intentions, without a push to go beyond surface-level gestures and address the deeper, systemic inequalities that persist, they are an inadequate solution. Inclusion perversely reinforces the power of dominant groups to recognize and “bestow humanity” upon the victims of data violences, transforming inclusivity into a celebrated collective win rather than a critique of on recognition, sorting, and othering in the first place (Hoffman, 2021).

3.2.9.4 Tactical PR Efforts Over Meaningful Change

Hoffmann's concerns extend to the superficial nature of many inclusion efforts, which often prioritize public relations over substantive change. Companies and institutions may adopt the language of inclusion and diversity without implementing the deep, systemic changes necessary to address underlying issues. This tokenistic approach not only fails to address the root causes of inequality but also perpetuates a cycle of empty promises and minimal progress. Meaningful change requires a commitment to dismantling existing power structures and creating genuinely equitable systems, rather than relying on tactical public relations efforts to appease critics and maintain the status quo.

The aforementioned Meta example as highlighted by Broussard (2023) referenced the 2014 move by the company, then Facebook, to allow users to self-identify their gender identities with more than 50 options. The underlying database, however, categorized users as male, female, or null, erasing any non-binary identity for the purposes of data analytics and targeted advertising (Broussard, 2023).

The recent retreat from Diversity, Equity, and Inclusion (DEI) commitments by major tech companies in 2023 underscores a critical disjunction between public promises and

actual practices (Elias, 2023). Corporations including Google and Meta publicly committed to substantial DEI initiatives after the civil unrest sparked by the murder of George Floyd in 2020, including increased representation of underrepresented groups. Recent reports, however, show that by mid-2023 many of these initiatives saw significant entrenchment (Elias, 2023). Alphabet, Google's parent company had pledged $12 million to organizations addressing racial inequities and $25 million in Google Ad Grants to organizations fighting racial injustice[5], while Facebook pledged $10 million[6]. Both tech giants have since reduced DEI staff, downsized programs, and significantly cut budgets for external DEI groups. For instance, Google's 2023 DEI budget cuts impacted initiatives such as the Early Career Immersion (ECI) program, which was designed to support underrepresented talent but saw no new cohort hired due to an uncertain hiring outlook. Similarly, Meta's Sourcer Development Program, aimed at diversifying corporate technology recruiting, was heavily scaled back (Elias, 2023).

As the tech industry increasingly intensifies its focus on AI, a reduction of diverse voices in the development of AI risks not only an amplification of existing biases and discrimination, but is poised to create new technologies that further perpetuate these inequalities. Tangible reductions in DEI efforts suggest a prioritization of immediate economic concerns over sustained equitable practices.

[5] Standing with the Black community, A message from our CEO, Google
[6] Mark Zuckerberg Facebook post, May 31, 2020

4. Methodology

4.1 Description of the Methodology

A mixed-method approach was applied combining qualitative methods to explore the study area. This approach includes a literature review, ethnographic interviews, and an online survey and data collection form. Through these methods, the aim was to generate comprehensive data, to inform an experimental project design, yielding both qualitative and quantitative results.

4.1.1 Literature Review

The foundation of this work is built on an extensive literature review, drawing from key scholarly sources in primary categories: social media and censorship (including algorithmic censorship, content moderation, and free speech), gender, AI, and discrimination (including gender and surveillance, algorithmic bias, and fairness), and critical studies on algorithms (including algorithmic impacts, ethics, and big data privacy). This literature review identified a research gap, motivating the design of a project that extends beyond academic discourse to propose new database structures. These structures are informed by a synthesis of literature, interviews, surveys, data collection methods, and established guiding principles. These principles will be discussed in a later section[7].

A comprehensive bibliography has been included.

4.1.2 Interviews

Ethnographic interviews were conducted with five participants via Zoom, Google Meet,

[7] See Experimentation, Prototype

and WhatsApp. The interviewees, who included a sex and intimacy educator, a sextech founder, a sex worker, a lingerie model and adult entertainment content creator, and a queer woman, shared their experiences with online censorship. These interviews provided rich, qualitative data focusing on individual experiences rather than large-scale trends. Participants were recruited through targeted outreach within the researcher's network and community groups. Per the request of two interviewees to remain anonymous, all respondents have been anonymized for simplicity and full transcripts have not been shared; anecdotes of their lived experiences and direct quotes from the interviewees have been included. All interview participants were offered $25 CAD as compensation for their time and contributions. To date, all participants have either not responded directly to the offer or declined to accept it. The interviews highlighted diverse perspectives on the impact of digital censorship and AI classification systems.

The data and the findings presented are by no means exhaustive and there are many gaps in the data including, but not limited to, geography, ethnicity, and sexuality. All interviewees were female presenting English speakers and were based in Canada, the United States, and Brazil, respectively. None of the interviewees offer a counter narrative of having either no experience with censorship or positive experiences with it, nor have narratives been included from those moderating the content.

4.1.3 Online Survey and Data Collection Form

An online survey distributed via Typeform received 59 responses, while a data collection form on Google Forms garnered one anonymous contribution to the dataset. The survey aimed to assess public understanding of AI systems' classification and representation of human bodies and to provide information for those looking to engage more deeply with the research.

When asked how familiar they were with the ways in which AI systems classify and represent the human body, 34.7 percent said they are not very familiar, another 30.6

percent said somewhat familiar, and 26.5 percent said not at all familiar. And yet, while only a small minority (8.2 percent) claimed to have a good understanding of how AI systems classify and represent the human body, 38.8 percent of all respondents suspect their bodies have been censored, suppressed, or misrepresented online by AI-powered platforms or applications.

Another key finding was 63.3 percent said they believe inherent biases and assumptions built into the algorithms are the reason why AI systems may classify, represent, or censor bodies in ways that are biased or exclusionary, while 59.2 percent feel it's due to societal norms and power structures that influence the design of these systems, indicating that within this sample group, there is a significant awareness of both the technical and sociocultural factors contributing to biased AI behavior. This dual recognition suggests that respondents are not only cognizant of the inherent limitations within AI algorithms but also define the broader context in which these systems operate. Such insights are crucial for informing further research and discussions on how to mitigate biases in AI, underscoring the need for interdisciplinary approaches that address both technological and societal dimensions of AI ethics and governance.

The survey questions and data collection questions have been included in the annex.

4.2 Description of the Methodology

The primary question driving this research is: How do social media platforms' censorship algorithms impact the representation and visibility of marginalized bodies, particularly in the context of gender and sexuality?

To thoroughly investigate this question, several secondary questions were addressed:

1. What are the underlying mechanisms and biases within social media platforms' censorship algorithms?

2. How do these algorithms affect the lived experiences and digital presence of marginalized communities?
3. What strategies can be developed to mitigate algorithmic bias and enhance equity and fair representation in digital spaces?

Objectives for Each Question

1. Understanding Mechanisms and Biases:
    a. Objective: To analyze the technical and sociocultural factors that contribute to algorithmic bias in social media platforms.
    b. Working Methods: Context frame rendering, literature review, and field observation.

2. Impact on Marginalized Communities:
    a. Objective: To document and synthesize the experiences of individuals from marginalized groups regarding censorship and misrepresentation by AI algorithms.
    b. Working Methods: Exploratory interviews and field observation.

3. Mitigation Strategies:
    a. Objective: To design interventions aimed at reducing heteronormative standards and promote more equitable representation.
    b. Working Methods: Applied design and prototype evaluation.

Working Methods

Exploratory Level Methods to Obtain Information

1. Context Frame Rendering Method:

- This method involved setting a comprehensive framework that contextualizes the study within broader sociotechnical systems.

2. Literature Review Method:
    - An extensive review of existing scholarly work provided a theoretical foundation and identified research gaps. Sources span areas such as algorithmic censorship, gender and AI discrimination, and the ethical implications of big data. This method helped in framing the research questions and understanding previous findings and methodologies.

3. Field Observation Method:
    - Direct observation of social media interactions and platform behaviors offered insights into how censorship algorithms function in real time. This method identified data on the frequency and nature of content removal or suppression, and its impact on users' digital experiences.

4. Exploratory Interviews:
    - Conducting semi-structured interviews with individuals from the sex-positive and queer communities provided qualitative data on personal experiences with censorship. This method focused on understanding the nuanced impacts of algorithmic decisions on users' online visibility and expression.

Generative Data or Solution Level Methods

1. Project Method in Applied Design:
    - This method involved the iterative design and development of interventions aimed at addressing the identified biases in censorship algorithms. It included the creation of a prototype based on the insights gained from exploratory methods and literature review. The prototype has not been tested.

5. Experimentation

This section details the design project executed during the Elisava Masters in Responsible AI, focusing on the development of the Deviant Dataset. This project aims to challenge and reconstruct the conventional methods of data collection and classification, incorporating principles of Design Justice, Privacy by Design, and Data Feminism, as well as the Feminist Data Set and Value Sensitive Design.

5.1 Presentation

The Deviant Dataset is a critical design tool and speculative project developed to explore alternative approaches to data classification and collection. Its primary goal is to challenge existing paradigms by questioning traditional data points and collection methods, while promoting a more contextual and temporal approach to data labeling. The project is participatory and community-driven, allowing contributors to manage and shape the dataset. This initiative aims to include voices traditionally excluded from data systems and design processes, positioning these individuals as experts and crucial contributors to the dataset's development. The project name, “Deviant Dataset,” is an act of reclamation, celebrating divergence from societal norms as a positive and empowering act.

> **deviant**[8]
> dē-vē-ənt
>
> adjective: deviant
> *straying or deviating especially from an accepted norm. "deviant behavior"*

[8] Deviant definition, Mirriam-Webster Dictionary

noun: deviant; plural noun: deviants
someone or something that deviates from a norm. Especially: a person who differs markedly (as in social adjustment or behavior) from what is considered normal or acceptable

5.2 Methodology

The design process for the Deviant Dataset follows a detailed methodology inspired by principles of Design Justice, Privacy by Design, Data Feminism, the Feminist Data Set, and Value Sensitive Design.

1. Context Frame Rendering Method: This initial phase involved understanding the broader context of data collection and classification systems. It examined historical and contemporary practices, highlighting the power structures that influence data management. With virtually all classification systems hidden in the so called “black box”, a thorough analysis was not possible of individual datasets, but a critical examination of current practices and the sociotechnical environments in which they operate was.

2. Literature Review Method: An extensive literature review was conducted to gather insights from scholars on algorithmic censorship, algorithmic bias, gender bias in AI, ethical data practices, and critical algorithm studies. This review provided the theoretical foundation for the project and highlighted the necessity for alternative data classification approaches.

3. Field Observation Method: Observations in communities and platforms where data exclusion occurs provided real-world insights. These observations helped synthesize the lived experiences of those affected by conventional data practices. These observations also helped inform the design by identifying pain points and opportunities for innovation.

4. Method of Exploratory Interviews: Ethnographic interviews with individuals affected by digital censorship were conducted. These interviews focused on personal narratives to uncover the nuances of digital platform interactions and AI classifications. These interviews offered qualitative insights into the experiences of those excluded by traditional data systems, emphasizing the importance of user-centered design.

5. Project Method in Applied Design: The core of the project involved creating the structure of the Deviant Dataset, a dataset framework that allows users to self-identify and manage their data inputs.

5.2.1 Guiding Principles

The development of the Deviant Dataset has been profoundly influenced by three guiding frameworks: the Design Justice Principles, Privacy by Design, and Data Feminism. These principles collectively inform the dataset's ethical foundation and operational choices, ensuring that it aligns with values of justice, privacy, and inclusivity. Secondary influences are the Feminist Data Set and Value Sensitive Design.

A full list of the three primary guiding frameworks and their principles can be found in the Annex.

Design Justice Network Principles

Firstly, the Design Justice Network Principles[9], established through the Allied Media Conference[10], emphasize the role of design in empowering communities and dismantling oppressive systems. These principles advocate for centering the voices of those directly impacted by design outcomes, prioritizing community needs over designer

[9] Design Justice Network Principles
[10] Allied Media Conference

intentions, and viewing change as an emergent process from collaborative and accessible practices. This aligns with the Deviant Dataset's anti-datafication approach, which allows users to define their own data points and categories, thereby sustaining their autonomy and agency. By acting as facilitators rather than experts, designers of the Deviant Dataset ensure that the system evolves in response to the lived experiences and contributions of its users, supporting more sustainable, community-led outcomes. Moreover, by honoring traditional, indigenous, and local knowledge, the dataset uplifts pre-existing community practices, reinforcing the principle of non-exploitative solutions that reconnect users to their communities and environments.

Some of the most influential principles include:

> Principle 1
> We use design to sustain, heal, and empower our communities, as well as to seek liberation from exploitative and oppressive systems.
>
> Principle 2
> We center the voices of those who are directly impacted by the outcomes of the design process.
>
> Principle 6
> We believe that everyone is an expert based on their own lived experience, and that we all have unique and brilliant contributions to bring to a design process.
>
> Principle 8
> We work towards sustainable, community-led and -controlled outcomes.

Privacy by Design

Secondly, the principles of Privacy by Design[11], conceived by Ann Cavoukian, underscore the importance of integrating privacy into the design and operation of systems from the outset. This proactive stance on privacy, where it is embedded by default, directly influences the Deviant Dataset's structure. The dataset's emphasis on user control over data expiration and privacy settings embodies the principle of proactive and preventive privacy measures, ensuring that privacy is the default setting and embedding it into the lifecycle design. Transparency and user-centric respect are intended through clear usage rights and full disclosure of any external partnerships, aligning with the Privacy by Design principles of visibility and transparency. This approach to privacy ensures that the system not only protects user data but also builds trust through openness and respect for user autonomy.

Some of the most influential principles include:

> Principle 1: Proactive not Reactive; Preventative not Remedial
> Anticipate, identify and prevent privacy invasive events before they occur.
>
> Principle 6: Visibility and Transparency — Keep it Open
> Assure stakeholders that privacy standards are open, transparent and subject to independent verification.
>
> Principle 7: Respect for User Privacy — Keep it User-Centric
> Protect the interests of users by offering strong privacy defaults, appropriate notice, and empowering user-friendly options.

[11] Privacy-Enhancing Technologies: The Path to Anonymity

Data Feminism

Lastly, Data Feminism[12], as articulated by Catherine D'Ignazio and Lauren Klein, focuses on examining and challenging power dynamics within data practices. This framework influences the Deviant Dataset by advocating for the elevation of emotion and embodiment, rethinking binaries and hierarchies, and embracing pluralism. The dataset's flexible, user-defined data inputs challenge traditional binary data structures and hierarchies, promoting a pluralistic approach to data representation. By considering the context in which data is created and used, the Deviant Dataset respects the diverse experiences of its users, making visible the often-invisible labor that goes into data generation and management. This alignment with Data Feminism principles ensures that the dataset not only collects data but does so in a manner that is just, equitable, and reflective of the diverse realities of its users.

Some of the most influential principles include:

> Principle 1: Examine power
> Data feminism begins by analyzing how power operates in the world.
>
> Principle 2: Challenge power
> Data feminism commits to challenging unequal power structures and working toward justice.
>
> Principle 4: Rethink binaries and hierarchies
> Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression.
>
> Principle 5: Embrace pluralism
> Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and

[12] Data Feminism, MIT Press

experiential ways of knowing.

Principle 6: Consider context
Data feminism asserts that data is not neutral or objective. It is the product of unequal social relations, and this context is essential for conducting accurate, ethical analysis.

Feminist Data Set

The Feminist Data Set, as conceived by Caroline Sinders[13], provides a framework that emphasizes the importance of ethical data collection and representation from a feminist perspective. This includes participatory methods to gather data and the deliberate and careful curation of data. This aligns with the Deviant Dataset's approach to focus on individual narratives through ethnographic interviews, rather than an emphasis on large scale trends or data collection, thereby promoting a more nuanced and accurate representation of lived experiences. It also advocates for transparent data practices and clear communication about how data is collected, used, and shared. This is mirrored in the Deviant Dataset's intention to share objectives, methodologies, and usage rights to contributors.

Value Sensitive Design

In their book, *Value Sensitive Design: Shaping Technology with Moral Imagination*, Batya Friedman and David G. Hendry outline Value Sensitive Design (VSD). VSD advocates for ethical consideration to be integrated at every stage of the technological design process. Value Sensitive Design begins with identifying the values that are to be embedded in the technology. For the Deviant Dataset, these values include inclusivity, respect for diversity, and user autonomy. By explicitly addressing these values, the project aims to create a dataset that reflects the complexities of human sexuality and identity, countering the reductive tendencies of traditional AI systems. Empirical

[13] Feminist Data Set, Caroline Sinders

investigations in VSD involve understanding how stakeholders experience the technology. The Deviant Dataset includes interviews and participatory research with individuals who have experienced censorship online. This stakeholder engagement ensures that the dataset is grounded in the real-world experiences of those it aims to represent, providing a more comprehensive and empathetic approach to data collection.

5.3 Prototype

The prototype of the Deviant Dataset is a participatory, crowd-sourced, and community-driven dataset, positioning the contributors as the ones that manage it. The Deviant Dataset's participatory structure has been chosen as the people of the Deviant Dataset are not just excluded from these systems, but also from the design process itself. Rather than making assumptions about the well being of those historically excluded, it is an attempt to center their voices and better understand the power structures at play that typically prevent their inclusion (Selbst, 2019). This is also inspired by the principles of Design Justice, as is the recognition that all humans have a valuable lived experience and these experiences and narratives have the right, and an imperative, to be shared. The deviants of the dataset are thus positioned as experts in their own right and crucial to the design process.

Unlike conventional systems that use fixed categories, the Deviant Dataset enables self-identifying inputs and optional categories, challenging the rigid structures of traditional data management. Individuals can label specific pieces of personal data as restricted or classified under safe mode–concepts typically imposed by platforms–prioritizing privacy preferences.

Data within the dataset is designed to be ephemeral, with users determining its expiration. This approach aligns with the European Union's General Data Protection Regulation's (GDPR[14]) right to be forgotten, offering greater control over personal information. The prototype's proposed continuous beta approach includes ongoing improvements and open communication channels to address vulnerabilities proactively through tactics like bug bounties, a system where companies or websites offer rewards through compensation to those who find bugs or other vulnerabilities within the system. The emphasis on proactive rather than reactive measures demonstrates a commitment to system integrity and user trust.

[14] EU GDPR

Human oversight, moderation, and curation is also integrated into the automated processes of the Deviant Dataset. Every automated check is supplemented by a human review, ensuring that the technology works collaboratively with properly trained and compensated humans. This dual-layer verification not only enhances accuracy but also imbues the system with a level of empathy and contextual understanding that purely automated systems often lack.

Its current structure, characterized by its emphasis on self-identifying inputs and optional categories, renders it non-machine readable and thus a more speculative project, a place for temporal identities and futures, rather than a technical one. However, this speculative framework aims to establish a new paradigm for data management, advocating for smaller scale, slower, and more ethical datasets. By prioritizing user autonomy and privacy over machine readability, the Deviant Dataset challenges conventional data practices, proposing a model that balances technological advancement with ethical considerations. This approach envisions a future where data systems are more attuned to individual rights and societal values, offering a viable alternative to the impersonal and often exploitative nature of large-scale datafication.

Key features of the prototype include:

- User-Defined Data Inputs:
    - Users can define their own data points, enhancing expression and reducing the reductive nature of categorization.
- Privacy and Control:
    - Users have the ability to impose restricted or safe modes on their data points, control data expiration, and ensure their privacy preferences are honored.
- Human Oversight:
    - The system integrates human oversight into automated processes to enhance empathy and contextual understanding.

5.4 Evaluation of the Results

The results of the Deviant Dataset project were analyzed both qualitatively and quantitatively:

- Qualitative Evaluation: Feedback from users and interviewees indicated a high level of satisfaction with the user-centered approach. The empowerment derived from self-identifying data inputs and the ability to control privacy settings were highlighted as significant benefits. However, challenges related to the non-machine readable nature of the dataset were noted, suggesting areas for future improvement.

- Quantitative Evaluation: Usage metrics, such as the number of contributions and user engagement levels, were tracked to assess the platform's adoption and effectiveness. Initial data indicated weak community participation, but a positive reception of the platform's principles. Survey conducted revealed a significant awareness of the technical and sociocultural factors influencing biased AI behavior, underscoring the platform's educational impact.

5.5 Limitations

The research presented here confronts the complexities of integrating AI with nuanced understandings of gender identities and body positivity. While strides have been made in exploring these intersections, several limitations were encountered during the research process that warrant scientific honesty and transparency.

One of the primary challenges was the limited access to comprehensive and diverse datasets. Despite efforts to obtain data, all repositories were either inaccessible due to privacy regulations or kept in the “black box” by default.

Time constraints posed another significant challenge. The scope of the research required extensive longitudinal studies to fully identify the long-term implications of AI on gender and body representation. However, the time available for experimentation, iteration, and evaluation of this project was insufficient to conduct such in-depth analysis. Consequently, some findings are preliminary and necessitate further validation through extended research.

An intended series of small workshops with affected communities did not materialize, representing a critical gap in the project. These iterative workshops should have been integral to the project's development, providing continuous feedback and ensuring the platform met the real needs of those experiencing censorship. The absence of this iterative process is seen as a crucial oversight, potentially limiting the platform's effectiveness and inclusivity.

Despite positive initial indicators, the lack of sustained and interactive engagement underscores the necessity of more proactive community-building efforts. Iterative workshops would not only enhance participation but also ensure that the platform evolves in line with users' needs and experiences. This approach is essential for the platform to achieve its full potential in supporting those affected by biased AI behavior and censorship.

Additionally, the technological infrastructure to support truly inclusive AI systems is still in its infancy. While theoretical frameworks and conceptual models have been proposed, practical implementation remains a significant hurdle. The scalability of these solutions and their integration into existing platforms pose technical and logistical challenges that were not fully resolved within the scope of this research.

The concept of “non-machine readable” data emerged as a potential solution to protect user privacy and prevent misuse of sensitive information, and yet is inherently non-interoperable with any other system or platform. The implementation of such data at scale requires further exploration. The current research did not fully address the mechanisms for storage, retrieval, and access control of data, which are critical for ensuring long-term sustainability and ethical management of these datasets.

Conclusions

The exploration of the intersection between virtual spaces, censorship, and the representation of marginalized bodies leads to several critical conclusions. Initially posed in the introduction, the central question concerns the impact of online censorship on sex-positive and marginalized bodies and how it influences societal perceptions of sexuality. The analysis confirms the initial hypothesis: that the design and implementation of database structures and classification systems within digital platforms inadequately reflect the diversity and complexity of human identities, thereby perpetuating existing societal biases and inequalities. These systems, entrenched in patriarchal, hyper-capitalist Western frameworks, are particularly resistant to accommodating marginalized identities, such as those of sex-positive and queer individuals.

The findings highlight that rigid classification systems, such as the binary gender model, fail to encompass the fluidity and spectrum of human identities. This inflexibility not only misrepresents nonbinary, queer, trans, intersex, and gender-nonconforming individuals but also perpetuates their marginalization by invalidating their experiences within computational infrastructures. The persistence of such outdated sociocultural norms in technical systems underscores a significant misalignment between social progress and technical implementation.

The study further reveals that algorithmic decision-making processes often reflect and reinforce societal prejudices. Algorithms used in content moderation, for instance, frequently misidentify or unfairly target marginalized bodies, leading to censorship and reduced visibility. This phenomenon is particularly evident in the experiences of sex-positive and queer bodies, which are disproportionately policed and silenced online. Such biases are not merely technical flaws but are deeply rooted in the broader sociocultural context, where multiple axes of power and oppression intersect (Collins, 1990).

The impact of digital platform policies, such as those stemming from legislative measures like FOSTA/SESTA, exacerbates the deplatforming and shadow banning of sex workers. These policies limit sex workers' access to essential online spaces for communication, support, and advocacy, reinforcing their stigmatization and isolation. The research underscores the broader challenges of designing inclusive and equitable sociotechnical systems that adapt to and reflect the complexities of human life.

On a personal level, the research underscores the broader societal implications and their significance. The experiences of individuals like DD1, an advocate for sex-positive education, discussed the challenges of disseminating essential information on sexual health and intimacy due to content restrictions and censorship. DD2, a plus-sized woman of color working as a lingerie model and content creator, revealed how algorithmic biases and societal standards collectively silence her voice, reduce her visibility, and hinder her ability to fully express and monetize her authentic self online. DD3, a partner and director of a sextech company, detailed the adverse effects of losing numerous social media accounts due to censorship, despite adhering to all guidelines, which severely limits her ability to educate and engage with her audience about sex in a comprehensive manner. DD4, a queer woman, articulated the psychological impact of algorithmic biases and societal standards on her sense of self and well-being. Lastly, DD5, detailed the adverse effects of deplatforming and shadow banning, which not only hinder her ability to work safely and communicate effectively but also reinforce her stigmatization and isolation. These narratives reveal how algorithmic biases and societal standards stifle creative expression, hinder a person's sense of sexuality and self-worth, and highlight the urgent need for more inclusive and supportive digital environments.

While the research underscores the significant challenges posed by entrenched societal structures in crafting ethical and equitable data systems and virtual spaces, it also highlights the importance of continued intervention and advocacy. The development of initiatives like the Deviant Dataset represents a hopeful blueprint for addressing these challenges, promoting more inclusive and supportive digital environments. This work

demonstrates that although the path to creating equitable digital spaces is fraught with obstacles, it is both necessary and possible to strive for systems that better reflect and support the complexities of human identities.

Contributions and Recommendations

This research makes contributions to understanding the complex intersections between AI, gender identities, and body and sex positivity. By highlighting the biases and limitations inherent in current AI technologies, it provides a foundation for developing more inclusive and equitable systems. The findings also underscore the importance of integrating more contextual and individualized data into AI models and the necessity of robust ethical frameworks to guide their development and deployment.

Contributions

1. Theoretical Insights: The research offers valuable theoretical insights into the pervasive biases in AI systems. It emphasizes the need for AI technologies to move beyond binary classifications and accommodate a broader spectrum of identities and bodies.

2. Framework Proposal: The study proposes a framework for designing AI systems that respect and celebrate diversity. This framework of the Deviant Dataset could serve as a blueprint for other researchers or developers aiming to create more inclusive digital environments.

3. Policy Implications: By identifying the gaps in current AI practices, the research informs policy development aimed at protecting marginalized communities. It advocates for policies that enforce the right to be forgotten and ensure ethical data management.

Recommendations

For researchers interested in furthering this field, several recommendations are put forth:

1. Access to Diverse Datasets: Future research should prioritize gaining access to more diverse and comprehensive datasets. Collaborating with organizations that manage such data and advocating for open data policies can help address this limitation. Researchers could also consider methods to create synthetic data that mirrors real-world diversity without compromising privacy.

2. Workshops with Affected Communities: To operationalize these recommendations, conducting workshops with those experiencing censorship, such as sex workers, LGBTQIA+ people, and other marginalized groups, is crucial. These workshops should focus on understanding their specific challenges and needs regarding AI and digital platforms. By providing hands-on training and open dialogues between developers, researchers, and those impacted by censorship, these workshops can help participants better define and implement the principles outlined in this research, ensuring that the voices and experiences of those directly impacted by AI biases are central to the development of more inclusive technologies.

3. Readable Non-Machine Readable Data: Further exploration is needed to implement non-machine readable data. Researchers should focus on developing efficient storage, retrieval, and access control mechanisms that balance data utility with privacy protection.

4. Interdisciplinary Collaboration: Advancing this research requires collaboration across disciplines, including computer science, gender studies, sociology, and ethics. Individuals from historically excluded communities should also be primary collaborators. Such interdisciplinary approaches can yield more holistic solutions and ensure that diverse perspectives are incorporated into AI development.

5. Design Thinking Applications: Applying research findings in a design thinking context can drive practical innovations. Researchers should explore how

inclusive design principles can be integrated into AI development processes and create tools to help practitioners analyze and improve their systems.

6. Longitudinal Studies: Conducting longitudinal studies is crucial to define the long-term impacts of AI on representation. Such studies can provide deeper insights into how AI systems evolve and their sustained effects on marginalized and historically excluded communities.

7. Ethical Frameworks and Policy Advocacy: Developing and advocating for ethical frameworks is essential. Researchers could engage with policymakers to create regulations that enforce fairness and inclusion in AI systems. This includes advocating for the right to be forgotten and ensuring transparent and accountable data management practices.

References and Bibliography

Broussard, M. (2023). More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech. Cambridge, MA. The MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, 81, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Butler, J. (2011, [1993]) Bodies That Matter: On the Discursive Limits of “Sex”. London and New York: Routledge.

Collins, P. (1990). Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment. Boston, MA. Hyman.

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press

Crawford, K., & Joler, V. (2018). Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources. AI Now Institute and Share Lab. Available at: http://anatomyof.ai

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*: Vol. 1989: Iss. 1, Article 8. Available at: http://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

Davisson, A., & Alati, K. (2024). “Difficult to Just Exist”: Social Media Platform Community Guidelines and the Free Speech Rights of Sex Workers. Social Media + Society, 10(1). https://doi.org/10.1177/20563051231224270

Elias, J. Tech companies like Google and Meta made cuts to DEI programs in 2023 after big promises in prior years [Internet]. CNBC. Published Dec 22, 2023. Available at: https://www.cnbc.com/2023/12/22/google-meta-other-tech-giants-cut-dei-programs-in-2023.html

Friedman, B., & Hendry, D. *Value Sensitive Design: Shaping Technology with Moral Imagination*. Mit Press, 2019.

Foucault, M. (1978) The History of Sexuality: An Introduction, vol. I. New York, NY. Pantheon Books.

Futrell, S. (2023). "Lip(s) Service: A Socioethical Overview of Social Media Platforms' Censorship Policies Regarding Consensual Sexual Content". Cybersecurity Undergraduate Research Showcase. https://digitalcommons.odu.edu/covacci-undergraduateresearch/2023fall/projects/6

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for Datasets. *Communications of the ACM*, *64*(12), 86-92. https://doi.org/10.48550/arXiv.1803.09010

Hoffmann, A. L. (2021). Terms of inclusion: Data, discourse, violence. *New Media & Society*, *23*(12), 3539-3556. https://doi.org/10.1177/1461444820958725

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Kahneman, D., Sibony, O., Sunstein, C. R. (2021). Noise: A Flaw in Human Judgment. New York, Little, Brown Spark.

Kantayya, S. (2020). *Coded Bias.* 7th Empire Media.

Katyal, S., Jung, J. (2021). The Gender Panopticon: Artificial Intelligence, Gender, and Design Justice. *UCLA Law Review*. https://doi.org/10.2139/ssrn.3760098

Keyes, O. (2018). The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact*. 2, CSCW, Article 88. https://doi.org/10.1145/3274357

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society,* 20(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. & Gebru, T. (2019). Model Cards for Model Reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229). https://doi.org/10.1145/3287560.3287596

Noble SU (2018) Algorithms of Oppression: How Search Engines Reinforce Racism. New York, NY. NYU Press.

Perez, S. X confirms plans for NSFW Communities [Internet]. Tech Crunch; Published Mar 29, 2024. Available at: https://techcrunch.com/2024/03/29/x-confirms-plans-for-nsfw-communities/

Riccio, P., Oliver, L., Escolano, F., Oliver, N. (2022). Algorithmic Censorship of Art: A Proposed Research Agenda. International Conference on Innovative Computing and Cloud Computing (ICCC). Available at: https://computationalcreativity.net/iccc22/papers/ICCC-2022_paper_128.pdf

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3). https://doi.org/10.1007/s42979-021-00592-x

Selbst, A,. Boyd, D., Friedler, S., Venkatasubramanian, S., Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 59-68. https://doi.org/10.1145/3287560.3287598

Sinders, C. (2017). The Feminist Data Set. [Project]. Funded by Clinic for Open Source Arts (COSA). Available at: https://carolinesinders.com/wp-content/uploads/2020/05/Feminist-Data-Set-Final-Draft-2020-0526.pdf

Snow, O. Why AI Is Detaining Sex Workers at the Border—and You May Be Next [Internet]. Daily Beast; Published Dec 2, 2023. Available at: https://www.thedailybeast.com/why-ai-is-detaining-sex-workers-at-the-borderand-you-may-be-next

Southerton, C., Marshall, D., Aggleton, P., Rasmussen, M. L., & Cover, R. (2021). Restricted modes: Social media, content classification and LGBTQ sexual citizenship. *New Media & Society,* 23(5), 920-938. https://doi.org/10.1177/1461444820904362

Stinson, C. (2022). Algorithms are not neutral: Bias in collaborative filtering. *AI and Ethics*, *2*(4), 763-770. https://doi.org/10.48550/arXiv.2105.01031

Stegeman, H. M., Are, C., & Poell, T. (2024). Strategic Invisibility: How Creators Manage the Risks and Constraints of Online Hyper(In)Visibility. *Social Media + Society, online first*, 1-12. https://doi.org/10.1177/20563051241244674

Testa, J. Etsy vs. Sex [Internet]. The New York Times; Published Jul 2, 2024. Available at: https://www.nytimes.com/2024/07/02/style/etsy-sex-toys-ban.html

Wang, S., & Spronk, R. (2023). “Big data see through you”: Sexual identifications in an age of algorithmic recommendation. *Big Data & Society*, 10(2). https://doi.org/10.1177/20539517231215358

Annex

A1 Survey Questionnaire

How familiar are you with the ways in which AI systems classify and represent the human body?
A Very familiar
B Somewhat familiar
C Not very familiar
D Not at all familiar

Are you aware of instances where your own digital body or the bodies of people you know have been censored, suppressed, or misrepresented by AI-powered platforms or applications?
A Yes, I am aware of specific instances
B I suspect it has happened, but I'm not sure
C No, I'm not aware of any such instances

Has anyone you know ever experienced having their body censored, misclassified or misrepresented by an AI system?
A Yes, this has happened to someone I know
B I suspect it has happened, but I'm not certain
C No, I'm not aware of any such instances
D I'm not sure

What do you think are some of the reasons why AI systems may classify, represent, or censor bodies in ways that are biased or exclusionary?
A Lack of diversity in the training data used to develop the AI models
B Inherent biases and assumptions built into the algorithms
C Profit-driven motives of the companies developing the AI

D Societal norms and power structures that influence the design of these systems
E I'm not sure

How interested would you be in learning more about the technical and social aspects of how AI systems represent and classify the human body?
A Very interested
B Somewhat interested
C Not very interested
D Not at all interested

What types of resources would you find most helpful in understanding these issues and how to challenge them?
A Case studies and examples of biased body representation in AI
B Workshops on dataset design and classification systems
C Webinars with researchers in the space
D Educational resources like PDFs, pamphlets, explainer documents
E Artistic projects, visual installations
F Other

What are some of the common ways that AI systems classify or categorize human bodies?
A By biological sex
B By perceived gender identity (man, woman, non-binary)
C By physical characteristics (height, weight, skin color, etc.)
D By ability status (able-bodied, disabled)
E I'm not sure

Why is it important to understand how AI systems classify and represent human bodies?
A To identify potential biases and limitations in how these systems perceive and categorize embodied identity

B To understand how these classifications can impact the experiences and opportunities of different individuals and communities
C To challenge the normative assumptions that often underlie AI-driven body representation
D I'm not sure

What are some potential consequences of having your body misclassified or misrepresented by an AI system?
A Denial of access or services
B Reinforcement of harmful stereotypes and biases
C Emotional distress and feelings of invalidation
D Perpetuation of exclusion and marginalization
E Increased surveillance
F I'm not sure

How important do you think it is to challenge and resist the normative ways in which AI systems classify and represent human bodies?
A Extremely important
B Somewhat important
C Not very important
D Not at all important
E I'm not sure

One of the key motivations behind The Deviant Dataset is to challenge the ways in which AI systems often classify and represent the human body. When an individual's embodied identity does not fit neatly into the normative categories encoded in a dataset, they are frequently seen as an outlier, an error, or a "deviant" case. This framing of non-conforming bodies as "deviant" can have serious consequences, leading to the exclusion, marginalization, and misrepresentation of individuals and communities whose lived experiences do not align with dominant societal norms. In light of this context, how do you understand the choice to name this project "The Deviant Dataset"?

A The name is a deliberate reclamation of the "deviant" label, empowering those whose bodies have been censored, suppressed, or marginalized
B The name highlights how current datasets and AI systems often treat non-normative bodies as "deviant" or problematic
C The name signals a commitment to challenging the very notion of what is considered "normal" or "acceptable" in terms of embodiment
D I'm still not sure about the reasoning behind the name choice

Would you be interested in contributing a submission to The Deviant Dataset?
A Yes, I would be interested in contributing
B Maybe, I'd like to learn more
C No, I'm not interested in participating

The next few slides will help me understand who has responded to this survey

Sex:
A Female
B Male
C Intersex
D Prefer not to say
E Other

Sexual identity:
A Asexual
B Aromantic
C Bisexual
D Pansexual
E Heterosexual
F Homosexual
G Queer
H Unsure

I Prefer not to say

J Other

Current geographic location:

A Africa

B Asia

C Australia

D Europe

E North America

F South America

A2 Data Collection Form

Self-Identification - short text answer

Let's get to know each other quickly first. In the space below, self-identify in a way that feels reflective of who you actually are. No check boxes or predefined answers. You are the expert of the parts of yourself that truly define you. This could be your sun, moon, or rising sign, gender, race, birth place, marital or parental status, neurodiversity, age, ethnicity, disability status, religion, occupation, or hobbies.

Story or Image Description - long answer

Provide a detailed description of an instance when your body was censored online. Explain, as best you can, why you think the algorithm or AI censored you, flagged the post, or suppressed it.

Data Expiration - multiple choice

How long do you think different data points that classify you remain accurate? Put a different way, if you could set an expiration date on your inclusion in a database, what would that be?
A 3-6 months
B 1-2 years
C 3-5 years
D 10 years
E Indefinitely
F No specific timeline but the ability to update or delete my data when I choose
G Other

Expiration Date - multiple choice

How long do you consent to your data being included here in the Deviant Dataset?
A 3-6 months
B 1-2 years
C 3-5 years
D 10 years
E Indefinitely
G Other

Permissions - checkboxes

Specify the permissions you grant for this data. Select all options that apply.
A I consent to my submission being included in the Deviant Dataset
B I consent to my submission being used for research purposes
C I consent to my submission being used in educational or academic contexts
D I consent to my submission being used for artistic or creative projects

Selfie - multiple choice

Would you like to contribute an image to the Deviant Database? This does not need to include your face or any identifying features. You can upload a section of your arm, an ear, an image of your hair. The "selfie" acts only as a visual representation of the bodies being censored.
A Yes
B No

Selfie - file upload

If you answered yes above, snap a shot and upload it here. Hit the Browse button once you click on File Upload and you will see the option to take a photo or upload something from your library.

Right to be forgotten - multiple choice

Thank you for participating! Regardless of your answers, all Deviant Dataset submissions have the right to be forgotten at any time. I understand that I can withdraw my consent and request the removal of my submission from The Deviant Dataset at any time by contacting the dataset's organizers at thedeviantdataset@gmail.com
A I understand

A3 Guiding Principles

A3.1 Design Justice Network Principles

1. We use design to sustain, heal, and empower our communities, as well as to seek liberation from exploitative and oppressive systems.
2. We center the voices of those who are directly impacted by the outcomes of the design process.
3. We prioritize design's impact on the community over the intentions of the designer.
4. We view change as emergent from an accountable, accessible, and collaborative process, rather than as a point at the end of a process.
5. We see the role of the designer as a facilitator rather than an expert.
6. We believe that everyone is an expert based on their own lived experience, and that we all have unique and brilliant contributions to bring to a design process.
7. We share design knowledge and tools with our communities.
8. We work towards sustainable, community-led and -controlled outcomes.
9. We work towards non-exploitative solutions that reconnect us to the earth and to each other.
10. Before seeking new design solutions, we look for what is already working at the community level. We honor and uplift traditional, indigenous, and local knowledge and practices.

A3.2 Privacy by Design

1. Proactive not Reactive; Preventative not Remedial
   Anticipate, identify and prevent privacy invasive events before they occur.

2. Privacy as the Default Setting
   Build in the maximum degree of privacy into the default settings for any system

or business practice. Doing so will keep a user's privacy intact, even if they choose to do nothing.

3. Privacy Embedded into Design
   Embed privacy settings into the design and architecture of information technology systems and business practices instead of implementing them after the fact as an add-on.

4. Full Functionality — Positive-Sum, not Zero-Sum
   Accommodate all legitimate interests and objectives in a positive-sum manner to create a balance between privacy and security because it is possible to have both.

5. End-to-End Security — Full Lifecycle Protection
   Embed strong security measures to the complete lifecycle of data to ensure secure management of the information from beginning to end.

6. Visibility and Transparency — Keep it Open
   Assure stakeholders that privacy standards are open, transparent and subject to independent verification.

7. Respect for User Privacy — Keep it User-Centric
   Protect the interests of users by offering strong privacy defaults, appropriate notice, and empowering user-friendly options.

A3.3 Data Feminism Principles

1. Examine power
   Data feminism begins by analyzing how power operates in the world.

2. Challenge power
   Data feminism commits to challenging unequal power structures and working toward justice.

3. Elevate emotion and embodiment.
   Data feminism teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world.

4. Rethink binaries and hierarchies
   Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression.

5. Embrace pluralism
   Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing.

6. Consider context
   Data feminism asserts that data is not neutral or objective. It is the product of unequal social relations, and this context is essential for conducting accurate, ethical analysis.

7. Make labor visible
   The work of data science, like all work in the world, is the work of many hands. Data feminism makes this labor visible so that it can be recognised and valued.

isec
lisboa

INSTITUTO
SUPERIOR
DE EDUCAÇÃO
E CIÊNCIAS

tgraf
isec lisboa

CENTRO INTERNACIONAL
DE ESTUDOS E INVESTIGAÇÃO
EM TECNOLOGIAS GRÁFICAS
E COMUNICAÇÃO CIENTÍFICA